contingency tables with fixed margins. *Comput. Statist. Data Anal.* **3** 159–185.

VOLLSET, S. E. and HIRJI, K. F. (1991). A microcomputer program for exact and asymptotic analysis of several 2×2 tables. *Epidemiology* **2** 217–220.

VOLLSET, S. E., HIRJI, K. F. and AFIFI, A. A. (1991). Evaluation of exact and asymptotic interval estimators in logistic analysis of matched case-control studies. *Biometrics* **47** 1311–1325.

VOLLSET, S. E., HIRJI, K. F. and ELASHOFF, R. M. (1991). Fast computation of exact confidence limits for the common odds ratio in a series of 2×2 tables. *J. Amer. Statist. Assoc.* **86** 404–409.

WHITE, A. A., LANDIS, J. R. and COOPER, M. M. (1982). A note on the equivalence of several marginal homogeneity test criteria for categorical data. *Internat. Statist. Rev.* **50** 27–34.

YATES, F. (1934). Contingency tables involving small numbers and the χ^2 test. *J. Roy. Statist. Soc. Supp.* **1** 217–235.

YATES, F. (1984). Tests of significance for 2×2 contingency tables. *J. Royal Statist. Soc. Ser. A* **147** 426–463.

ZELEN, M. (1971). The analysis of several 2×2 contingency tables. *Biometrika* **58** 129–137.

ZELEN, M. (1972). Exact significance tests for contingency tables embedded in a 2^n classification. *Proc. Sixth Berkeley Symp. Math. Statist. Probab.* **1** 737–757. Univ. California Press, Berkeley.

ZELTERMAN, D. (1987). Goodness-of-fit tests for large sparse multinomial distributions. *J. Amer. Statist. Assoc.* **82** 624–629.

Comment

Edward J. Bedrick and Joe R. Hill

We congratulate Professor Agresti for his comprehensive review of exact inference with categorical data. We share his enthusiasm for exact conditional methods and believe that the coming years will produce many important computational breakthroughs in this area.

The mechanics of conditioning on sufficient statistics to generate reference distributions for estimation, testing and model checking with loglinear models for Poisson data and logistic regression models for binomial data are well-known, but the utility of conditioning in these settings is not universally agreed upon. Furthermore, the role of conditioning in the analysis of discrete generalized linear models with noncanonical link functions has received little attention from most of the statistical community. As a result, scientists and statisticians are familiar with conditional methods, but many are unsure how such methods should be incorporated into an overall strategy for analyzing categorical data. We feel that the use and abuse of conditional methods will not be fully understood or appreciated without such a strategy. We hope that Professor Agresti's survey and the ensuing discussions stimulate further work in this direction.

Edward J. Bedrick is an Assistant Professor, Department of Mathematics and Statistics, University of New Mexico, Albuquerque, New Mexico 87131. Joe R. Hill is an R&D Specialist at EDS Research, 5951 Jefferson Street, NE, Albuquerque, New Mexico 87109.

CHECKING LOGISTIC REGRESSION MODELS

We would like to convey some of our recent work on model checking for logistic regression and some of our thoughts regarding conditional inference. For the sake of simplicity, we assume that a single model is under consideration. A little notation is required. The usual logistic regression model has two distinct parts: a sampling component and a structural component. The sampling component specifies that $Y = (Y_1, \ldots, Y_n)'$ is a vector of independent binomial random variables with $Y_i \sim \mathrm{Bin}(m_i, \pi_i)$. The structural or regression component of the model is given by

$$\text{(1)} \qquad \operatorname{logit}(\pi) = X\beta,$$

where $\operatorname{logit}(\pi)$ is an $n \times 1$ vector of log-odds with elements $\log\{\pi_i/(1-\pi_i)\}$, X is an $n \times p$ full-column rank design matrix with ith row x_i', and β is a $p \times 1$ vector of unknown regression parameters. Under model (1), $S = X'Y$ is sufficient for β. Let $\hat{\pi}$ be the MLE of π under this model.

The distribution of the data $\mathrm{pr}(Y; \beta)$, indexed by β, can be factored into the marginal distribution of the sufficient statistic S, and the conditional distribution of the data given the sufficient statistic:

$$\mathrm{pr}(Y; \beta) = \mathrm{pr}(Y \mid S)\mathrm{pr}(S; \beta).$$

Taking a Fisherian point of view (Fisher, 1950), inferences about β are based on $\mathrm{pr}(S; \beta)$, whereas model checks use the conditional distribution $\mathrm{pr}(Y \mid S)$. Letting $s_{obs} = X' y_{obs}$ be the observed value of the sufficient statistic for the logistic model, the reference distribution for model checking is

$$\text{(2)} \qquad \text{pr}(Y = y \mid S = s_{obs}) = C_{obs} \prod_{i=1}^{n} \binom{m_i}{y_i}$$

where

$$C_{obs}^{-1} = \sum_{(y_1^*, \ldots, y_n^*)' \in \mathscr{Y}_{obs}} \prod_{i=1}^{n} \binom{m_i}{y_i^*},$$

$$\mathscr{Y}_{obs} = \{ y^* = (y_1^*, \ldots, y_n^*)',\ y_i^* \text{ an integer: } 0 \le y_i^* \le m_i \text{ and } X' y^* = s_{obs} \}.$$

Note that $\mathscr{Y}_{obs}$ is the set of response vectors that give the same value of the sufficient statistic as the observed data.

Model checking examines the adequacy of the postulated model by contrasting it with the observed data. Either the entire model (i.e., both the sampling and structural components) or specific features of the model may be checked, neither requiring reference to alternative models. A model check is implemented by specifying a test statistic $t(Y)$ that quantifies the type of deviation of interest, and then computing a significance level for the statistic using the conditional reference distribution $\text{pr}(Y \mid S = s_{obs})$. Model checks are pure significance tests. Unlike a hypothesis test that requires a specified alternative, the results of a model check are provisional and only meant to provide insight into whether difficulties might arise if the model were to be used for inference. Of course, a series of model checks may suggest that the model needs to be revised.

We believe that conditioning on the sufficient statistic S is the only logically defendable approach to generating a (frequentist) reference distribution for checking logistic models. This view is controversial among frequentists, many of whom argue that conditional methods are needlessly conservative and that unconditional assessments are preferable. Indeed, this criticism has been leveled repeatedly at Fisher's exact test. It is our opinion that most of the criticisms are flawed because they are based on power comparisons, which are not relevant in significance tests. Fisher's exact test is a global significance test of the adequacy of the following model: $Y_i \sim$ independent $\text{Bin}(m_i, \theta)$, $i = 1, 2$. It is not a test of equal probabilities *assuming independent binomial responses*. Any mathematical connection between the two is largely coincidental and does not justify power comparisons of the exact test with hypothesis tests. Professor Agresti's linking of Fisher's exact test to a hypothesis test is unfortunate, as this will not help abate the confusion on this issue.

Fisher's exact test is, in fact, a simple case of a global check for a logistic regression model where the structural component of the model includes only an intercept. The probability function $t(Y) = \text{pr}(Y \mid S = s_{obs})$ serves as the test statistic for a global model check. The deviance and Pearson statistics are possible alternatives to $t(Y)$.

Bedrick and Hill (1990, 1991) discussed a general strategy for checking logistic models, including global model checks, outlier detection and goodness-of-link checks. Exact reference distributions for residuals can also be obtained in this framework. To illustrate the idea, let $e_i = (Y_i - m_i\hat{\pi}_i)/\{(1 - h_i)m_i\hat{\pi}_i(1 - \hat{\pi}_i)\}^{1/2}$ be the ith standardized Pearson residual, where h_i is the estimated case leverage. The e_i's are usually treated as standard normal variables, which is unrealistic if the sample sizes are small, say 5 or less. The inherent discreteness of the Y_i can be handled directly by conditioning on S and calibrating the e_i relative to the marginal conditional distribution $\text{pr}(Y_i \mid S = s_{\text{obs}})$. For example, when $m_i = 1$, the marginal conditional distribution of Y_i is Bernoulli with parameter $\text{pr}(Y_i = 1 \mid S = s_{obs})$. The conditional distribution of e_i is a location-scale change of the conditional distribution of Y_i because the fitted values and leverages are functions of s_{obs}. The conditional distribution of Y_i can be extremely skewed, multimodal or restricted to a proper subset of $[0, m_i]$. We prefer to assess the extremeness of the Y_i directly when the conditional reference distribution is not unimodal or not centered near the fitted value. Large differences between the normal and conditional assessments of the residuals are common in the presence of extreme skewness and multimodality. Although it is difficult to reconcile differences between conditional and unconditional assessments, we are uncomfortable with the normal approximation when the conditional reference distribution exhibits these characteristics.

The vasoconstriction data set analyzed by Pregibon (1981) highlights several issues concerning conditional versus unconditional asymptotic methods. Pregibon fitted a model with two predictors to the 39 binary responses. Normal approximations to the deviance and Pearson residuals suggest that observations 4 and 18 in Pregibon's data listing are poorly accounted for by the model. Although no statistician would take literally these large sample assessments, how much faith can one place in benchmarking residuals with a normal distribution, even if it is used solely as a diagnostic to point out poorly fitted points? We computed $\mathscr{Y}_{obs}$, after rounding off the covariate values to three decimal places. The reference set $\mathscr{Y}_{obs}$ contains three response vectors (see Table 1), each with conditional

TABLE 1
Time in seconds needed to generate $\mathscr{Y}_{obs}$ on SUN SPARCstation IPC

Source	p	n	$\sum_i m_i$	$\# \mathscr{Y}_{obs}{}^a$	Time
Bedrick and Hill (1990)	2	10	274	1637	14.5
Bedrick and Hill (1990)	6	23	53	6034	3.6
Cook and Weisberg (1982, page 193)	3	30	33	131	1.6
Finney (1947, page 47)	2	6	283	1496	2.3
Pregibon (1981)	3	39	39	3	127.1

[a] $\# \mathscr{Y}_{obs}$ is the number of responses in $\mathscr{Y}_{obs}$.

probability 1/3 under the model. Clearly, in this situation, there is insufficient information to criticize the model based on conditional assessments. The conditional and the unconditional asymptotic assessments of the residuals are in conflict, in part, because of the near degeneracy of the conditional reference distribution. We agree with Professor Agresti that many frequentist statisticians would be uncomfortable with the conditional approach here. However, in the light of the inadequacies of the unconditional asymptotic approximations, what direction should one turn to handle such data? Are bootstrap methods a viable alternative to asymptotics for model checking? Some researchers have suggested so (see Williams, 1987), but this issue has not been fully explored. In closing this section, we would like to ask Professor Agresti whether he has some good advice on general strategies for handling such problems?

AN ALGORITHM TO GENERATE $\mathscr{Y}_{obs}$

Model checking is feasible only when the reference set $\mathscr{Y}_{obs}$ that determines $\mathrm{pr}(Y \mid S = s_{obs})$ can be generated readily. A complete enumeration of $\mathscr{Y}_{obs}$ is necessary because a variety of assessments are usually required. We will discuss an algorithm we developed to generate $\mathscr{Y}_{obs}$. For simplicity, assume that the model contains an intercept and, then without loss of generality, that the covariates are nonnegative. The ideas extend easily to arbitrary covariates values.

For models with only an intercept term, $\mathscr{Y}_{obs}$ is equivalent to the set of $2 \times n$ contingency tables with fixed-column totals $m_1, \ldots, m_n$, and fixed-row totals $s_{obs} = \sum_i y_{i,obs}$ and $\sum_i m_i - s_{obs}$. As is well-known, the enumeration of these tables can require enormous CPU time even for tables with modestly sized margins. The computations of $\mathscr{Y}_{obs}$ for models that also include covariates may appear more difficult. In fact, fewer computations are necessary because each covariate imposes an additional constraint on the set of feasible responses. For models with several covariates, $\mathscr{Y}_{obs}$ can often be obtained in a small fraction of the time necessary to generate the set of $2 \times n$ contingency tables with fixed margins.

We generate $\mathscr{Y}_{obs}$ by recursively looping between the minimum and maximum potential values for $y_1, \ldots, y_n$, and checking whether the generated responses are in $\mathscr{Y}_{obs}$. The minimum and maximum values are determined from the counts assigned to preceding cells in the recursion and by "infeasibility criteria" (Hirji, Mehta and Patel, 1987) that restrict cell values based on the constraints imposed by the covariates. The infeasibility criteria eliminate potential responses from consideration once it becomes obvious that further steps in the recursion will not produce an element in $\mathscr{Y}_{obs}$.

The basic idea of the algorithm is straightforward. Given $Y_j = y_j$, $j = 1, \ldots, i-1$, define the minimum and maximum potential values of Y_i as follows: $\min_i$ is the maximum of 0 and the smallest integer Y_i such that

$$Y_i x_i \geq s_{obs} - \sum_{j=0}^{i-1} y_j x_j - \sum_{j=i+1}^{n} m_j x_j,$$

whereas $\max_i$ is the minimum of m_i and the largest integer Y_i such that

$$Y_i x_i \leq s_{obs} - \sum_{j=0}^{i-1} y_j x_j.$$

Here, the inequalities must be satisfied elementwise, and $y_0 = 0$, $x_0 = 0_{p \times 1}$. Because the covariates are nonnegative, any element in $\mathscr{Y}_{obs}$ with $Y_j = y_j$, $j = 1, \ldots, i-1$ must have $\min_i \leq Y_i \leq \max_i$. Potential responses in $\mathscr{Y}_{obs}$ are generated by looping between $\min_i$ and $\max_i$ $(i = 1, \ldots, n-p)$ to get $y_1, \ldots, y_{n-p}$, then solving for $y_{n-p+1}, \ldots, y_n$ by forcing the constraint $X'y = s_{obs}$:

$$\sum_{k=n-p+1}^{n} y_k x_k = s_{obs} - \sum_{j=0}^{n-p} y_j x_j.$$

The generated tail values $y_{n-p+1}, \ldots, y_n$ are not necessarily integers, so we must check whether $(y_1, \ldots, y_n)' \in \mathscr{Y}_{obs}$. The efficiency of our algorithm depends critically on the data labeling. The data should be arranged to minimize the amount of looping. The bounds $\min_i$ and $\max_i$ only crudely utilize that $\sum_i y_i$ is fixed by the inclusion of an intercept in the model. A more careful analysis typically provides tighter bounds, which results in faster execution time.

We implemented a FORTRAN version of our algorithm on a SUN SPARC station IPC with 16 Mb RAM. Table 1 gives the CPU seconds needed to generate $\mathcal{Y}_{obs}$ for several published data sets. The single covariate data sets ($p = 2$) are dose-response problems with relatively large sample sizes per dose, whereas the remaining examples have many small-sized samples. The summaries suggest that generating the reference set $\mathcal{Y}_{obs}$ is feasible for many practical problems. We are currently investigating alternative methods to widen the range of problems that can be handled. We hope to report these results in the near future.

A COMMENT ON NONCANONICAL LINK FUNCTIONS

In Section 2.2, Professor Agresti states:

> It is not possible to construct 'exact' confidence intervals for association measures that are not functions of the odds ratio. They do not occur as parameters in generalized linear models with Poisson or binomial random component using canonical links. Thus, the usual conditioning arguments do not eliminate nuisance parameters.

Professor Agresti's comment raises several important issues concerning conditional inference and model checking that we will try to clarify. To be precise, we will consider three distinct models for the two-by-two table. Let Y_1 and Y_2 be independent and, for $i = 1, 2$, let Y_i be a binomial random variable with index m_i and natural parameter α_i, that is,

$$\begin{aligned}\log\{\mathrm{pr}(Y = y; \alpha)\} = {} & y_1\alpha_1 + m_1\log\{1 + \exp(\alpha_1)\} \\ & + y_2\alpha_2 + m_2\log\{1 + \exp(\alpha_2)\} \\ & + c(m_1, m_2, y_1, y_2),\end{aligned}$$

which is a two parameter linear exponential family indexed by $\alpha = (\alpha_1, \alpha_2)$.

If

$$\alpha_1(\phi) = \log(\psi_0) + \log(\phi)$$

and

$$\alpha_2(\phi) = \log(\phi),$$

$0 < \phi < \infty$, for some fixed ψ_0, $1 \le \psi_0 < \infty$, then $S = Y_1 + Y_2$ is sufficient for ϕ, so $\mathrm{pr}(Y_1 = y_1 \mid S = s_{obs})$ does not depend on ϕ. (It does depend on ψ_0, which is known.) This, of course, leads to Fisher's exact test.

If, however,

$$\alpha_1(\pi) = \log\left\{\frac{\pi + (1 - \pi)\theta_0}{(1 - \pi)(1 - \theta_0)}\right\}$$

and

$$\alpha_2(\pi) = \log\{\pi/(1 - \pi)\},$$

$0 < \pi < 1$, for some fixed θ_0, $0 < \theta_0 < 1$, then $S = (Y_1, Y_2)$ is minimally sufficient for π. According to Basu (1979), this disallows Fisher's conditioning argument. Basu commented:

> Is there a logically compelling reason why we should reparameterize the model in terms of (ψ, ϕ) instead of (θ, π)? I cannot regard the Fisher conditionality argument as anything but an ad hoc method that appears to succeed once in a while but fails completely when the same problem is restated in a slightly different form.

Should Fisher's disciples be disturbed by this counterexample? The answer is a categorical *NO*. The two models previously described are not different parametrizations of the same model. They are fundamentally different models. In particular, the first model is a one-parameter linear exponential family. On the other hand, the second model is a one-parameter curved exponential family nonlinearly embedded in a two-dimensional linear model. These facts, not the failure of Fisher's conditioning argument, explain why the dimension of the minimal sufficient statistic is one for the first model and two for the second.

A careful Fisherian analysis of the second model uses the ancillary part of the minimally sufficient statistic $S = Y$ for model checking. This ancillary statistic is related to the curvature of the second model. Assuming the model passes inspection, inference about π must be made conditional on the ancillary (Hinkley, 1980).

If, instead of either of the these two models, $\alpha = (\alpha_1, \alpha_2)$ were unconstrained, then, in our opinion, no completely coherent argument has been given to justify conditioning on any statistic, no matter what the parameter of interest is. If the parameter of interest is the odds-ratio $\psi = \exp(\alpha_1 - \alpha_2)$, then conditioning provides a mathematical device to derive similar confidence intervals for ψ. On the other hand, if the parameter of interest is $\theta = (\pi_1 - \pi_2)/(1 - \pi_1)$, where $\pi_i = 1/\{1 + \exp(-\alpha_i)\}$, then this device does not work but reasonable confidence intervals still exist. Given that this mathematical device works only in special cases, is there any reason that conditional coverage should be desired?

In summary, for the first model, pr($Y \mid S$) is the reference distribution for model checking and pr($S; \phi$) is the reference distribution for inference about the parameter ϕ. For the second model, pr(A), where A represents the ancillary component of $S = Y$, is the reference distribution for model checking and pr($S \mid A; \pi$) is the reference distribution for inference about the parameter π. For the third model, model checking is not possible, and pr($Y; \alpha$) is the reference distribution for inference about any parameter of interest (i.e., any function of α).

Comment

Diana E. Duffy

1. INTRODUCTION

Professor Agresti is to be congratulated for a well-written and timely survey on exact conditional inference for contingency tables. At this point in time, 8 to 10 years after some of the key initial advances in computing strategies (Pagano and Halvorsen, 1981; Mehta and Patel, 1983; Pagano and Tritchler, 1983a, 1983b), it is instructive to take stock of both where the field is presently and where the field may be headed. For practitioners and applied statisticians, Agresti offers a practical introduction to currently available exact methods. For researchers in methodology and in statistical computing and algorithms. Agresti offers directions for possible future research.

Exact conditional inference for contingency tables involves assessing the exact (discrete) sampling distribution of test statistics and parameter estimates of interest after conditioning on the sufficient statistics for nuisance parameters. The sufficient statistics for nuisance parameters correspond directly to certain margins in the corresponding contingency table; as long as one operates within the arena of loglinear models, conditioning on these margins will eliminate the nuisance parameters. The exact sampling distribution of interest is then the distribution over all possible tables that could be observed with certain fixed margins (i.e., those margins fixed by the sampling design plus those margins fixed by the conditioning). I will refer to this set of tables as the conditional reference set. It is worth noting the following correspondence between conditional and unconditional problems: the conditional reference set for a problem with a set S_1 of margins fixed by the sampling design and a set S_2 of margins fixed by conditioning is identical to the sample space for a (different) problem in which margins in both S_1 and S_2 are fixed by the sampling design. For example, the conditional reference set for testing independence in a 2×2 table under product binomial sampling (one-fixed margin) is equivalent to the sample space for a 2×2 table with both margins fixed.

In this commentary, I would like to expand on two areas that offer challenges for future work. Throughout this discussion, I adopt Agresti's notation as described in his Section 1.2 in toto, and I refer to points in his paper by simply giving his name and the section number.

Diane E. Duffy is Director, Statistics and Data Analysis Research Group, Bellcore, 445 South Street, Morristown, New Jersey 07962-1910.

2. BAYES AND RELATED INFERENCES

The existing literature on Bayesian methods for analyzing contingency tables dates at least to Lindley (1964). One way to categorize the proposed methods is through the choice of prior. The simplest methods are those for 2×2 tables under product binomial sampling with beta priors; see Altham (1969, 1971) for examples. Generalizations to full multinomial sampling and to $I \times J$ tables for I or $J > 2$ lead to Dirichlet priors on the cell probabilities. These are discussed in Good (1967, 1975, 1976), Good and Crook (1974), Gunel and Dickey (1974), Crook and Good (1980), and Albert and Gupta (1982, 1983a, 1983b). Normal priors on logarithmic functions of the cell probabilities are discussed in Leonard (1975) and Nazaret (1987). Empirical Bayes analogs of the Dirichlet and normal approaches are described in Albert (1987) and Laird