COMMENTARIES

Commentaries are informative essays dealing with viewpoints of statistical practice, statistical education, and other topics considered to be of general interest to the broad readership of *The American Statistician*. Commentaries are similar in spirit to Letters to the Editor, but they involve longer discussions of background, issues, and perspectives. All commentaries will be refereed for their merit and compatibility with these criteria.

The Many Faces of Logistic Regression

DAVID STRAUSS*

Logistic regression has found wide acceptance as a model for the dependence of a binary response variable on a vector of explanatory variables. It can also be used, however, as a maximization algorithm for fitting a variety of other parametric models. The easy availability of logistic regression in standard packages is a major advantage; further, the regression diagnostics routinely supplied are frequently useful, even though the model being fitted is not logistic.

In some cases the objective function maximized is a likelihood, but the method seems to arise especially often in the maximization of a so-called pseudolikelihood. Applications include models from choice theory, spatial modeling, random graph theory, and educational testing.

KEY WORDS: Maximum likelihood; Pseudolikelihood estimation.

1. INTRODUCTION

Suppose we have a binary variable Y and want to model its dependence on a vector $\mathbf{x}$ of p explanatory variables by

$$E(Y) = P(Y = 1) = g(\beta'\mathbf{x}), \tag{1.1}$$

where β is a p vector of parameters. A common choice for $g(t)$ is

$$g(t) = \exp(t)/\{1 + \exp(t)\}, \tag{1.2}$$

the inverse of the standard logistic distribution function. In this case (1.1) can be written

$$\text{logit}\,\{P(Y = 1|\mathbf{x})\} = \boldsymbol{\beta}'\mathbf{x}, \tag{1.3}$$

where $\text{logit}(t) \equiv \log\{t/(1 - t)\}$. Equation (1.3) is a *logistic regression* model.

The applications of logistic regression to be discussed here differ from the above, in that a parametric model of some sort is specified and is to be fitted by maximization of an objective function. The latter may be the likelihood function, or in some cases a so-called pseudolikelihood (to be defined shortly). In a surprising variety of problems the maximization is formally equivalent to maximum likelihood for a suitably defined logistic regression model. The wide availability of computer packages to implement the latter is a major advantage. Moreover, even though the real model being fitted is not logistic, in most cases at least some of the regression diagnostics supplied by the package will be useful in assessment of the fit.

Some of the work on these topics is currently unpublished, and what is available is specialized to particular models and is scattered in the literature. The aim of this article is to give a unified account of applications in a variety of situations, to show some common features, and to increase awareness of these rather useful ideas.

In all the examples to be considered, formal maximum likelihood estimation for the logistic regression, assuming independent cases, is appropriate (even if the outcomes are in fact not independent). Implementation in computer packages is most commonly performed by iteratively reweighted least squares (McCullagh and Nelder 1983, sec. 2.5). Other fitting methods for logistic regression (such as the computationally simpler minimum χ^2 procedure) will not be exactly equivalent to maximization of the relevant objective function.

The next section gives some examples of how logistic regression may be used to implement maximum likelihood for other models. Section 3 summarizes the pseudolikelihood method and gives examples of how logistic regression can be used to maximize the pseudolikelihood.

2. MAXIMUM LIKELIHOOD BY LOGISTIC REGRESSION

2.1 Bradley–Terry Model

Our first example of parameter estimation by logistic regression is the well-known Bradley–Terry model for paired comparisons. In different settings the pairs of "stimuli" being compared might be psychophysical, such as sounds or lights of different intensities, consumer items, such as cars or soft drinks, or competing athletes.

*David Strauss is Professor, Department of Statistics, University of California at Riverside, Riverside, CA 92521. The author thanks Paul Holland for some helpful discussions and the referees and an associate editor for suggesting some substantial improvements to the article.

It is assumed that each comparison will result in a unique choice of one stimulus over the other. This might correspond to a judgment that one sound is the louder of the two, that one car is preferred to the other, or that one team beats the other. The various paired comparisons are assumed to be made independently. See Bradley (1985) for a summary, and Davidson and Farquhar (1976) for an extensive bibliography.

According to the Bradley–Terry model, for each of the p stimuli there is a parameter π_i such that

$$P(i > j) = \pi_i/(\pi_i + \pi_j), \quad 1 \le i, j \le p, \quad (2.1)$$

where $i > j$ means that stimulus i is chosen over j. A side condition, such as $\Sigma\pi_i = 1$, is evidently required. For the model to be identifiable it is also assumed that the pairs being compared are properly "connected," so that there is at least one sequence of comparisons linking any two stimuli in the set.

Currently, perhaps the most common way of fitting the model involves a rather ingenious use of loglinear modeling (see Fienberg 1977, p. 150 for the details). This method has the desirable features of providing standard errors of estimates and goodness-of-fit tests. An alternative approach, and one that shares these features, is to note that (2.1) can be written

$$P(i > j) = \text{expit}(\beta_i - \beta_j). \quad (2.2)$$

Here $\beta_i = \log \pi_i$ and expit is a convenient notation for the inverse of the logit function: $\text{expit}(t) = \exp(t)/\{1 + \exp(t)\}$. Equivalently,

$$\begin{aligned} \text{logit}\,\{P(i > j)\} &= \beta_i - \beta_j, \\ &= \mathbf{x}'\boldsymbol{\beta}, \end{aligned} \quad (2.3)$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ and $x_k = 1$ if $k = i$, -1 if $k = j$, and 0 otherwise. The likelihood function is the product of expression (2.2) over all paired comparisons; its maximization is thus equivalent to a maximum likelihood solution for the logistic regression model (2.3). Here the dummy p vector $\mathbf{x}$ serves as the explanatory variable, or "data," in the regression. As in (2.1), a side condition on the β's, such as $\Sigma\beta_i = 0$, is required. The logistic regression should be constrained to have no intercept; this is an option in all major packages.

An example of this is the analysis of 1987 American League baseball wins and losses, given by Agresti (1990, p. 371). He also shows how to fit and test an extension to the model that includes a parameter for the home team advantage. In addition, the residual diagnostics provided by the logistic regression program allow easy identification of any pairs of teams whose scores against each other are poorly predicted by the model.

2.2 Luce's Choice Model

The Bradley–Terry model can be written in the following way: there are parameters π_i such that when a set S_r of two stimuli is presented, the chance that stimulus i is selected is given by

$$P_i(S_r) = \pi_i \Big/ \sum \pi_j, \quad (2.4)$$

the sum being taken over j in S_r. A natural generalization of the model is to the case where (2.4) holds for sets S_r of any size. This is (almost) equivalent to the Choice Axiom of Luce (1959), which essentially states that the ratio of the choice probabilities of two stimuli is independent of alternatives in the choice set. Luce's model (2.4) is the point of departure for much of the probabilistic choice theory developed by mathematical psychologists; for a summary see, for example, Strauss (1985).

In numerical applications of the choice model (Luce 1977), a common procedure is maximum likelihood implemented by a general-purpose maximization routine. Apart from the typical problems of maximization in high-dimensional parameter spaces, this approach will supply few if any of the desired diagnostics statistics. Instead, however, the problem may be cast into the form of a *polytomous* logistic regression, as follows.

Suppose that the response variable for the ith comparison, y_i, is categorical, with values indexed by $k = 1, \ldots, c$. The polytomous logistic regression model is (Fienberg 1977, sec. 6.5)

$$P(y_i = k) = \exp(\boldsymbol{\gamma}_k'\mathbf{x}_i) \Big/ \sum_m \exp(\boldsymbol{\gamma}_m'\mathbf{x}_i). \quad (2.5)$$

The parameters here are the p vectors $\boldsymbol{\gamma}_k$, $k = 1, \ldots, c$. One choice of side condition to ensure identifiability is to set $\boldsymbol{\gamma}_1 = 0$. To express the Luce model (2.4) in this form, take $c = p$ and let $\mathbf{x}_i$ again be the dummy p vector whose ith component is 1 and the others 0. One also needs to be able to restrict the sum in (2.5) to the specified choice sets S_r; the polytomous logistic regression program in SYSTAT is one that allows this option.

2.3 Heterogeneous Poisson Process

This is a stochastic process that generates events (points) in time or in a spatial region of arbitrary dimension. It may be characterized by two properties (Diggle 1983, p. 52):

1. The numbers of events in disjoint regions are independent.
2. The probability of an event in a small region of area c, located at x, is given by

$$c\lambda(x) + o(c), \quad (2.6)$$

where $\lambda(x)$ is the *intensity function* of the process. The ordinary (homogeneous) Poisson process arises if $\lambda(x)$ is a constant. In many cases, however, it may be desirable to model a dependence of the intensity function on a vector of covariates $z(x)$. For example, the events may be the occurrence of trees in a spatial region. The covariates $\{z_j(x)\}$ might then include soil quality, water availability, and so forth, as well as the coordinates of x so as to model a polynomial trend surface.

Since the intensity $\lambda(x)$ is constrained to be nonnegative, a natural model for the dependence is

$$\lambda(x) = \exp\{\boldsymbol{\beta}'\mathbf{z}(x)\}, \quad (2.7)$$

where $\boldsymbol{\beta}$ is a vector of unknown parameters. This model (in a temporal context) has previously been considered by Mathers (1984) and many others. Mathers's method of analysis approximates the distributions of the number of events in various subregions, these being Poisson with means equal to the integral of (2.6) over their subregions, followed by maximization of the likelihood function.

A simpler estimation method begins by placing a fine grid over the region, with cells of size c, say. Let y_i be 1 if the ith cell contains an event and zero otherwise, and let $\mathbf{z}_i$ be the vector of covariates for the ith cell. According to (2.6) and (2.7),

$$\text{logit}\,\{P(Y_i = 1)\} = \ln c + \boldsymbol{\beta}'\mathbf{z}_i \qquad (2.8)$$

apart from an error of magnitude $o(c)$. Further, the Y_i are independent. Thus, apart from the error $o(c)$, maximum likelihood estimation of $\boldsymbol{\beta}$ is accomplished by a logistic regression of the y's on the $\mathbf{z}$'s. If one component of $\mathbf{z}$ is a dummy variable then the $(\ln c)$ term in (2.8) should be subtracted from the corresponding parameter estimate; otherwise the regression should be forced to have an intercept of $\ln c$.

The choice of c is discussed briefly in Section 3.3; the issue is that if c is too large, the error term $o(c)$ becomes serious, while if it is very small, the number of cells becomes large and the analysis expensive.

3. LOGISTIC REGRESSION FOR MAXIMIZATION OF PSEUDOLIKELIHOOD

3.1 Pseudolikelihood Estimation

Let $\mathbf{u} = (u_1, \ldots, u_s)$ have a likelihood function L indexed by a (possibly vector) parameter θ. It sometimes happens that maximum likelihood estimation of θ is difficult because L involves a complicated normalizing function of θ. For this reason (or for others, as we shall see) it may be advantageous to replace L by a *pseudolikelihood function*, a term and idea originating with Besag (1975) (see also Lindsay 1988). The construction is as follows.

Choose suitable functions $v_1, v_2, \ldots$ of $\mathbf{u}$; in applications the v's may consist of one or more components of $\mathbf{u}$, sums of components, and so on. Form conditional density functions $f_{ij}(v_i|v_j)$, assuming these to exist, and take the product of a collection of such densities. (Unconditional densities are allowed: one of the v's may be set to a constant.) The choice of terms in the product is a matter of convenience; we shall see examples shortly. Note that any normalizing constant in the likelihood cancels out in the conditional densities. Given a random sample $\mathbf{u}^{(k)}$, $k = 1, 2, \ldots$, from the distribution, one defines a pseudolikelihood (PL) for the data as any quantity of the form

$$\text{PL} = \prod_k \left[\prod_{ij} f_{ij}\{v_i^{(k)}|v_j^{(k)}\} \right]. \qquad (3.1)$$

By maximizing (3.1) one obtains a *maximum pseudolikelihood estimator* (MPE). Under regularity conditions the MPE can be shown to be consistent and asymptotically normally distributed, with asymptotic variance given by certain information-type quantities (see Arnold and Strauss 1991b). In general the MPE will not be a function of minimal sufficient statistics, and thus will not be fully efficient; but the inefficiency will often be slight, and compensated by dramatic computational simplifications.

In a considerable number of cases, maximization of the pseudolikelihood function can be conveniently implemented by a logistic regression. One instance, that of a network or graph model with Markovian dependence, is discussed in detail in a recent article (Strauss and Ikeda 1990). The balance of this section describes three other applications.

3.2 Lattice Models

This case appears to be the first where pseudolikelihood idea was used. The simplest form of lattice model is the celebrated *Ising model* (Ising 1925) of statistical mechanics, currently enjoying some statistical prominence as a prior distribution in work on image enhancement (Besag 1986). The model specifies a joint distribution for a rectangular array of binary variables y_{ij}. The *sites* (i, j) and (k, l) are said to be *neighbors* if either $i = k$ and $|j - l| = 1$ or $j = l$ and $|i - k| = 1$. Let S be $\Sigma\Sigma y_{ij}$, the number of sites with value 1, and let n_{ij} be the sum of y_{kl} over the four neighboring sites of (i, j). Write $N = (1/2)\Sigma\Sigma n_{ij}$. According to the Ising model, the probability of a realization $\mathbf{y}$ of the set of lattice variables $\{y_{ij}\}$ is given by

$$P(\mathbf{y}) = \{1/Z(\alpha, \beta)\} \exp(\alpha S + \beta N). \qquad (3.2)$$

The parameter β measures the intensity of the interaction; when β is zero the y_{ij} are Bernoulli with probability expit(α), while positive values of β promote clustering of like values of the y_{ij}. For example, the odds on the event $y_{ij} = 1$ increase by $\exp(\beta)$ for a unit increase in n_{ij}. The normalizing constant $Z(\alpha, \beta)$, known as the partition function, is notoriously intractable and the source of much anguish in statistical mechanics. Note, on the other hand, the simple form taken by the conditional probabilities:

$$P(y_{ij} = 1|\text{ all the other } y\text{'s}) = \text{expit}(\alpha + \beta n_{ij}). \qquad (3.3)$$

This led Besag (1975, 1977) to define a pseudolikelihood as the product of (3.3) over all i, j and to estimate α, β by its maximization. The consistency of this MPE does not follow from the result quoted in Section 3.1, but has been proved by Geman and Graffigne (1987).

Since, from (3.3),

$$\text{logit}\,\{P(y_{ij} = 1|\text{ all other } y\text{'s})\} = \alpha + \beta n_{ij}, \qquad (3.4)$$

it follows as before that the MPE is obtained by a formal maximum likelihood estimation of the logistic regression model. The data entries are simply the "response variables" y_{ij} and the "predictor variables" $(1, n_{ij})$; α and β are then the unknown parameters in the linear regression. Odencrantz (1988) presented one of the first applications of this idea. Exact calculation of the efficiency of the estimator seems not feasible, but sim-

ulation studies by Odencrantz (1988) and Ripley (private communication) suggest that the MPE loses little efficiency, compared to the maximum likelihood estimator, provided β is below the so-called *critical point*. Above that value the MPE may perform notably worse than the maximum likelihood estimator. [This last implication may be unimportant in statistical applications, since for large lattices with β above the critical point the realizations will contain infinite patches of zeroes and similar patches of ones. See Pickard (1987) for a discussion of critical phenomena in a statistical context.]

Estimation by linear logistic regression is easily adapted to various generalizations of the Ising model. For example, a term γm_{ij} may be added to the right side of (3.4), where m_{ij} is the sum of y_{kl} over the four "diagonally adjacent" neighbors of (i, j), and γ is the corresponding intensity parameter. That is,

$$\text{logit}\,\{P(y_{ij} = 1 | \text{all other } y\text{'s}) = \alpha + \beta n_{ij} + \gamma m_{ij}. \quad (3.5)$$

It is also easy to define generalizations of (3.2) to the *colored* lattice, where the y's are now c-valued polytomous variables (Strauss 1977). Such models may be appropriate for arrays of plants of differing species or with various possible states of health; see also Section 3.3. Particular patterns of interaction between different colors are expressible through constraints on the parameter matrix in a polytomous logistic regression.

As an illustration, Figure 1 gives data from Bartlett (1971) on the presence or absence of the plant *Carex arenaria* over a spatial region divided into a 24 × 24 lattice. Visual inspection, or some simple analyses, indicates clustering of the plants, and it may be of interest to fit some dependence models. A stepwise logistic regression (BMDP's PLR program) was used, with two predictor variables, the n_{ij}'s and the m_{ij}'s. The cells on the boundary of the lattice were excluded, as the predictor variables are undefined for them.

Table 1 gives various (pseudo)likelihood statistics that may help one assess the models. The first row corresponds to a model where the y_{ij}'s are independent, with $p_{ij} = P(Y_{ij} = 1)$ equal to a constant. Since each of the two regressor variables ranges from 0 to 4, there are 5 × 5 = 25 possible distinct patterns of covariates, of which 23 actually occur here. The saturated alternative model estimates each of the 23 p_{ij}'s by the observed proportion. Thus the deviance, or goodness-of-fit χ^2 statistic, 66.06, has 23 − 1 = 22 df. [Note that for this comparison we have a genuine likelihood ratio, so the usual χ^2 test is valid; in the other cases we are dealing with pseudolikelihoods, and the χ^2 reference values should only be used as a rough guide. An additional *caveat* applies to the nominal p values in Table 2: Of the 46 cells (0 or 1 response for each of 23 distinct covariate patterns), 21 have observed frequencies less than 5. For the comparison of a given pair of models, a more conservative treatment would be as follows: First pool all cells where the expected frequency according to *either* model is lower than some acceptable threshold (e.g., 3 or 5) and then compute the likelihood ratio and its p value for the reduced set. (The usual additivity property of deviances would then no longer apply.)]

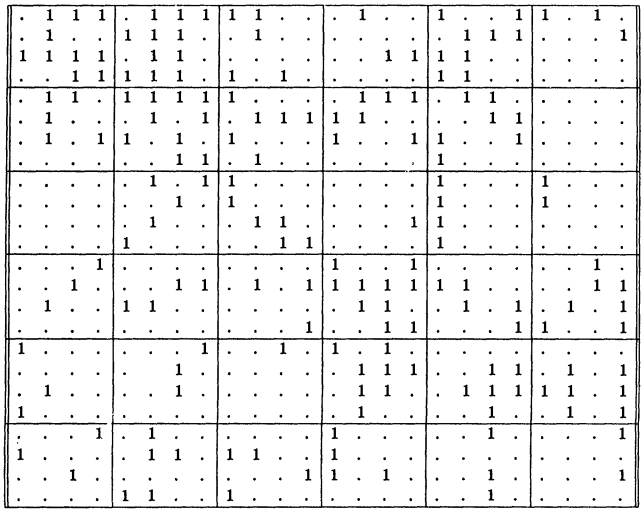

Figure 1. 24 × 24 Grid of Presence/Absence of the Plant Carex Arenaria. (From Bartlett 1971.)

The first variable added is the set $\{n_{ij}\}$, corresponding to the pure Ising model (3.2) or (3.4). The addition of a single parameter accounts for about half of the above deviance. The fit is still not ideal, as indicated by the model's deviance of 34.56 on 21 df relative to the saturated model. (A valid likelihood ratio test of this is not available; a Monte Carlo test, however, has been developed by Besag and Clifford 1989.) Addition of the "diagonal neighbors" count $\{m_{ij}\}$ seems to make an appreciable improvement. A rather loose interpretation is that the influence of neighboring cells is not fully

Table 1. Log-Pseudolikelihoods and Their Differences for Some Models of the Data of Table 1

Model	Parameters fitted	Log of (pseudo-) likelihood	Deviance, relative to saturated model	df	Nominal p value (based on χ^2 for deviance)	Deviance relative to previous model	df	Nominal p value
1. Independence	One: the common cell probability	−296.33	66.06	22	.00	—	—	—
2. Ising (3.4)	Two: α and β	−280.58	34.56	21	.03	31.50	1	.00
3. (3.5)	Three: α, β, and γ	−276.20	25.80	20	.17	8.77	1	.00
4. Saturated	23: all cell probabilities	−263.30	—	—	—	—	—	—

captured in the four immediately surrounding values. One could, of course, go on to include more distant neighbors than the diagonally adjacent ones.

The estimate of the intensity parameter β, 0.54, is the increase in log-odds on occupancy of a cell when n_{ij} is increased by 1. The estimate, not surprisingly, turns out to be intermediate between the two "coding method" estimates, 0.48 and 0.59, given by Besag (1972). (The coding method, an early example of pseudolikelihood, treats the lattice as a 24 × 24 black-and-white chessboard; one maximizes the likelihood of the values at the black squares conditional on those at the white squares, and then vice versa.) For the model that also includes the $\{m_{ij}\}$, the parameter estimates are

$$\beta = .45, \quad \gamma = .31$$

A log-odds interpretation for these is similar to the above. Just as in ordinary regression, the reduction in the coefficient β when an additional, positively correlated, variable is included is to be expected.

Diagnostics analogous to those mentioned in Section 2 are again available. They are omitted, as they seem not to provide any new insights in this example.

3.3 Spatial Interaction Models

Spatial point processes arise in many contexts; examples are the patterns of towns on a map, plants in a field, or gas molecules in a container. In these examples, and many others, the events (points) may be expected to "interact"; for example, the presence of a plant at a given location is likely to inhibit the presence of other plants in its immediate neighborhood (competition effects). At somewhat larger separations the plant may appear to attract others. An important class of models for such patterns of interacting points is that of *pair-potential models*, or *Gibbs distributions* (Diggle 1983, sec. 4.9). The class is unusual among spatial models in that an explicit joint distribution can be written down.

To obtain a Gibbs distribution for points in a (typically two-dimensional) domain D, one begins with a unit Poisson process on D. That is, the number of points n follows a Poisson distribution with mean 1, and given n we place the points independently and uniformly in D. The sample space Ω is the set of collections x of n points in D, for all $n \geq 0$. One then defines a *potential function* $U(x)$ on Ω to specify the interactions between pairs of points in x. It is usual to take

$$U(x) = \Sigma\, u(r_{ij}),$$

where r_{ij} is the distance between the ith and jth points in x, and u is a suitable *pair-interaction function*. The simplest nontrivial example is (Strauss 1975)

$$\begin{aligned} u(r) &= \beta \text{ if } r \leq r_0 \\ &= 0 \text{ if } r > r_0, \end{aligned} \tag{3.6}$$

where r_0 is the interaction range and the interaction parameter β must be nonpositive. Then U is simply β times the number of r_0-close pairs. A Gibbs distribution on Ω is determined by a "density" function

$$f(x) = (1/Z)\exp\{\alpha n + U(x)\}, \tag{3.7}$$

where α is an intensity parameter and Z is again the partition function. The function $f(x)$ specifies how likely the configuration x is, relative to the unit Poisson process. The latter corresponds to the case $\alpha = 1$ and $U \equiv 0$. Equation (3.6) gives a model of repulsion between pairs of r_0-close points.

In general, parameters of (3.7) cannot be estimated through maximum likelihood because (as usual) Z is intractable. Several estimation procedures (e.g., Ogata and Tanemura 1984) have been based on approximations to Z, but they are not consistent and tend to involve elaborate computations.

Following Besag (1977) one may place a fine grid over the spatial pattern, just as in Section 2.3. The cells C_i of the grid have common size c. Let n_i be the number of points in C_i. If c is taken to be suitably small, $P(n_i > 1)$ is negligible, and we have a binary lattice model. This differs from the Ising model in that each cell has many neighbors; nevertheless, the conditional probabilities are easily written down. With (3.6), for example, we have

$$\begin{aligned} &P(n_i = 1 | \text{all other } n_j\text{'s}) \\ &\qquad = \operatorname{expit}(\alpha + \beta \Delta U_i) + o(c) \end{aligned} \tag{3.8}$$

where ΔU_i is the number of points that are r_0-close to C_i. Besag's idea was to define a pseudolikelihood as the product of (3.8) over all i, and to estimate parameters by its maximization.

Once again, the method can conveniently be implemented by logistic regression. A number of desirable properties, such as convergence of the sequence of estimators as $c \to 0$, and consistency as both $c \to 0$ and the domain becomes large, can be shown to hold (Clyde and Strauss 1991). Their simulation studies suggest again that the method is about as good as maximum likelihood, in the few cases where the latter is feasible, at least for parameter values below the critical point (compare the comments in Section 3.2). Other simulations given in the paper deal with the bias corresponding to the $o(c)$ term. It appears that the bias is slight provided that c is chosen to be small enough that at most a few cells contain more than one event.

The logistic regression applies equally to a wide variety of models where the quantity ΔU in (3.8) is linear in the parameters. Just as in the lattice case, a natural generalization of the model is to the polytomous problem (e.g., species of several types, with differing patterns of interaction); as before, this can be handled with a polytomous logistic regression. One may also allow the density for a point at a given location in D to depend not only on the interactions, but also on exogenous variables, or on the local "fertility," exactly as in Section 2.3. One needs only to replace the term αn in (3.7) by $\Sigma\, \lambda(\mathbf{z}_i)$, the sum being taken over all n locations indexed by the $\mathbf{z}_i$, for a suitably chosen intensity function λ. Logistic regression still provides a consistent estimator of unknown parameters in the λ function.

The models described in the previous subsections are all exponential families involving intractable normalizing constants. We have seen that similar estimation methods apply to all three models. The models share many other features, such as feasible simulation methods and "stability" properties (Strauss 1986). Pseudolikelihood estimation is generally worth consideration in exponential families with awkward normalizing constants (Arnold and Strauss 1991a).

3.4 Rasch Model

Our final example comes from the theory of mental testing. Suppose a group of subjects attempt to answer a set of J items, and let x_{ij} be an indicator variable for the event that subject i gives the correct answer to item j. A widely used model for the distribution of the x_{ij} is that these variables are independent with

$$P(x_{ij} = 1) = \text{expit}\{a(\theta_i - \beta_j)\}. \tag{3.9}$$

Here θ_i represents the ability of the ith subject, β_j the difficulty of the jth item, and a is a scale constant that may (and will be) set to 1 here. Evidently a side condition on each of θ and β is required for identifiability. Equation (3.9) represents the simplest form of the Rasch model.

By taking logits in (3.9) we obtain a logistic regression scheme for maximum likelihood estimation. In practice this will generally not be satisfactory, as the number of subjects is often very large and the abilities θ_i are typically regarded as nuisance parameters. We now show one way in which the pseudolikelihood approach can be used to eliminate the nuisance parameters.

Take as pseudolikelihood the product

$$\prod_i \left[\prod_{j<k} P(x_{ij} | x_{ij} + x_{ik}) \right], \tag{3.10}$$

which is of the form (3.1). Had we taken $j \neq k$ instead of $j < k$ in (3.10), it would come to the same thing. When $x_{ij} + x_{ik} = 0$ or 2, the corresponding conditional probability in (3.10) is sure to be 1, and the term may be omitted from the product. Hence we only need consider "discordant" cases in (3.10), that is, triples i, j, k where subject i gets one of items j, k right. Now it follows from (3.9), with a little algebra, that

$$P(x_{ij} = 1 | x_{ij} + x_{ik} = 1) = \text{expit}(\beta_k - \beta_j). \tag{3.11}$$

The important point here is that dependence on the θ's has been eliminated. Once again, maximization of the pseudolikelihood (3.10) can be carried out with a logistic regression: for each item pair j, k the "trials" are the subjects given discordant answers and the "successes" are the cases where the correctly answered item is the jth. As mentioned, a side condition on the β's is needed, and the constant term in the logistic regression should be suppressed.

The efficiency of the estimator is in general very complicated algebraically, but an explicit form can be given in the simple case where all θ's are equal and there are only two items. It can then be shown that the MPE is fully efficient iff $\beta_1 = \beta_2$. In educational testing practice, when there are usually many subjects, the loss of efficiency may be rather unimportant.

If this logistic regression scheme looks familiar, it is not surprising: we have come all the way back to our first example. For a pair of discordant items creates a "choice" as to which is correctly answered, and we have seen that the probabilities for such choices in the Rasch model satisfy the Bradley–Terry model. The pseudolikelihood estimator for the former is equivalent to maximum likelihood applied to the latter.

4. CONCLUSION

We have seen that a variety of models, seemingly quite unrelated to logistic regression, may be fitted to data with the aid of that technique. Many other applications are possible; for example, another pseudolikelihood estimation scheme for the Rasch model is obtained if one conditions on the sum of more than two item variables. The logistic regression is then polytomous, and would be expected to be more efficient than the method given, at the price of greater complexity.

When logistic regression is available it has several attractive features. To summarize:

1. It avoids the use of general-purpose maximization routines, which sometimes suffer from convergence problems.
2. It can be performed conveniently with familiar computer packages.
3. Just as in standard logistic regression, a large number of candidate models can conveniently be compared. For a fairly extensive example, from social network analysis, see Strauss and Ikeda (1990).
4. Useful diagnostics are readily available. When the quantity maximized is a genuine likelihood, the usual likelihood ratios and asymptotic standard errors of parameters are directly applicable. Even when a pseudolikelihood is used, the pseudolikelihood ratios provide an informal basis for model selection. Another basis, available as an option in many packages, is the number of correct classifications of the binary variable, based on the fitted regression with an optimal cutpoint (as in discriminant analysis). Plots of observed and expected proportions may be used to indicate unusual cells.

[*Received September 1989. Revised January 1992.*]

REFERENCES

Agresti, A (1990), *Categorical Data Analysis*, New York: John Wiley.

Arnold, B. C., and Strauss, D. J. (1991a), "Bivariate Distributions With Conditionals in Prescribed Exponential Families," *Journal of the Royal Statistical Society*, Ser. B, 53, 365–375.

——— (1991b), "Pseudolikelihood Estimation: Some Examples," *Sankyā*, 53, 233–243.

Bartlett, M. S. (1971), "Two-Dimensional Nearest Neighbor Systems and Their Ecological Applications," in *Statistical Ecology* (Vol. I), University Park, PA: Pennsylvania State University Press, pp. 179–194.

Besag, J. E. (1972), "On the Statistical Analysis of Nearest-Neighbor Systems," European Meeting of Statisticians, Budapest.

——— (1975), "Statistical Analysis of Non-Lattice Data," *The Statistician*, 24, 179–195.
——— (1977), "Some method of Statistical Analysis of Spatial Data," *Bulletin of the International Statistical Association*, 47, 77–92.
——— (1986), "The Statistical Analysis of Dirty Pictures" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 48, 259–302.
Besag, J. E., and Clifford, P. (1989), "Generalized Monte Carlo Significance Tests," *Biometrika*, 76, 633–642.
Bradley, R. A. (1985), "Paired Comparisons," in *Encyclopedia of Statistical Sciences* (Vol. 6), eds. S. Kotz and N. L. Johnson, New York: John Wiley, pp.
Clyde, M. C., and Strauss D. J. (1991), "Logistic Regression for Spatial Pair-Potential Models," in *Spatial Statistics and Imaging: Proceedings of a 1988 AMS-IMS-SIAM Joint Summer Research Conference, Bowdoin College, Maine*, ed. A. Possolo, Lecture Notes—Monograph Series, Hayward, CA: Institute of Mathematical Statistics, pp. 14–30.
Davidson, R. R., and Farquhar, P. H. (1976), "A Bibliography on the Method of Paired Comparisons," *Biometrics*, 32, 241–252.
Diggle, P. J. (1983), "*Statistical Analysis of Spatial Point Patterns*," New York: Academic Press.
Fienberg, S. E. (1977), "*The Analysis of Cross-Classified Data*," Cambridge, MA: MIT Press.
Frank, O., and Strauss, D, J. (1986), "Markov Graphs," *Journal of the American Statistical Association*, 81, 32–42.
Geman, S., and Graffigne, C. (1987), "Markov Random Field Image Models and Their Application to Computer Vision," in *Proceedings of the International Congress of Mathematics, 1986*, ed. A. M. Gleason, Providence: American Mathematical Society.
Ising, E. (1925), "Beitrag Zur Theorie des Ferromagnetismus," *Zeitschrift fur Physik*, 31, 253–258.
Lindsay, B. G. (1988), "Composite Likelihood," *Contemporary Mathematics*, 80, 221–239.
Luce, R. D. (1959), "*Individual Choice Behavior*," New York: John Wiley.
——— (1977), "The Choice Axiom Twenty Years Later," *Journal of Mathematical Psychology*, 15, 215–233.
Mathers, C. D. (1984), "Maximum Likelihood Estimation of Exponential Polynomial Rate for Poisson Data," *Biometrical Journal*, 26, 33–38.
McCullagh, P., and Nelder, J. A. (1983), *Generalized Linear Models*, London: Chapman and Hall.
Odencrantz, J. (1988), "Parametric Inference for Binary and Colored Lattices," unpublished Ph.D. dissertation, University of California, Riverside, Dept. of Statistics.
Ogata, Y., and Tanemura, M. (1984), "Likelihood Analysis of Spatial Point Patterns," *Journal of the Royal Statistical Society*, Ser. B, 46, 496–518.
Pickard, D. K. (1987), "Inference for Discrete Markov Fields: the Simplest Nontrivial Case," *Journal of the American Statistical Association*, 82, 90–96.
Strauss, D. J. (1975), "A Model for Clustering," *Biometrika*, 62, 467–475.
——— (1977), "Clustering on Colored Lattices," *Journal of Applied Probability*, 14, 135–143.
——— (1985), "Luce's Choice Axiom and Its Generalizations," in *Encyclopedia of Statistical Sciences* (Vol. 5), eds. S. Kotz and N. L. Johnson, New York: John Wiley, pp.
——— (1986), "On a General Class of Models for Interaction," *SIAM Review*, 28, 513–527.
Strauss, D. J., and Ikeda, M. (1990), "Pseudolikelihood for Social Networks," *Journal of the American Statistical Association*, 85, 204–212.

On Generalized Score Tests

DENNIS D. BOOS*

Generalizations of Rao's score test are receiving increased attention, especially in the econometrics and biostatistics literature. These generalizations are able to account for certain model inadequacies or lack of knowledge by use of empirical variance estimates. This article shows how the various forms of the generalized test statistic arise from Taylor expansion of the estimating equations. The general estimating equations structure unifies a variety of applications and helps suggest new areas of application.

KEY WORDS: Composite null hypothesis; Empirical variance; Estimating equations; Information sandwich; Lagrange multiplier test; Misspecified likelihood; Observed information; Robust inference.

1. INTRODUCTION

Rao (1948) introduced score statistics having the form

$$S(\tilde{\theta})^T \tilde{I}_f^{-1} S(\tilde{\theta}), \qquad (1)$$

where $S(\theta)$ is the vector of partial derivatives of the log likelihood function, $\tilde{\theta}$ is the vector of restricted maximum likelihood estimates under H_0, and $\tilde{I}_f$ is the Fisher information of the sample evaluated at $\tilde{\theta}$. These test statistics are attractive because they only require computation of the null estimates $\tilde{\theta}$ and are asymptotically equivalent to Wald and likelihood ratio statistics under both null and Pitman alternative hypotheses (Serfling 1980, p. 156). In fact many common tests statistics such as the Pearson chi-square are score statistics or are closely related. A parallel development of (1) was begun by Aitchison and Silvey (1958) under the name "Lagrange multiplier statistic," and the econometrics literature uses this latter term. Introductions to score and Lagrange multiplier tests may be found in Breusch and Pagan (1980), Buse (1982), Engle (1984), Hosking (1983), and Tarone (1988).

The purpose of this note is to discuss the use of score tests in the general estimating equations situation where $\hat{\theta}$ is obtained by solving the vector equation $S(\theta) = 0$. Estimation methods which give rise to different $S(\theta)$ include maximum likelihood, least squares, robust M-estimation, and quasi-likelihood. In analogy with maximum likelihood estimation where $S(\theta) = \partial Q(\theta)/\partial\theta$ for the log likelihood $Q(\theta)$, I will call S the "score function"

*Dennis D. Boos is Professor, Statistics Department, North Carolina State University, Raleigh, NC 27695. The author thanks Ron Gallant and Len Stefanski for helpful discussions during the preparation of this article.