Title
Maximum Likelihood, Profile Likelihood, and Penalized Likelihood: A Primer

Journal
American Journal of Epidemiology, 179(2)

ISSN
0002-9262

Authors
Cole, Stephen R
Chu, Haitao
Greenland, Sander

Publication Date
2014-01-15

DOI
10.1093/aje/kwt245

Peer reviewed

American Journal of Epidemiology

Vol. 179, No. 2
DOI: 10.1093/aje/kwt245
Advance Access publication: October 29, 2013

Practice of Epidemiology

Maximum Likelihood, Profile Likelihood, and Penalized Likelihood: A Primer

Stephen R. Cole*, Haitao Chu, and Sander Greenland

* Correspondence to Dr. Stephen R. Cole, Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Campus Box 7435, Chapel Hill, NC 27599-7435 (e-mail: cole@unc.edu).

Initially submitted February 20, 2013; accepted for publication September 16, 2013.

The method of maximum likelihood is widely used in epidemiology, yet many epidemiologists receive little or no education in the conceptual underpinnings of the approach. Here we provide a primer on maximum likelihood and some important extensions which have proven useful in epidemiologic research, and which reveal connections between maximum likelihood and Bayesian methods. For a given data set and probability model, maximum likelihood finds values of the model parameters that give the observed data the highest probability. As with all inferential statistical methods, maximum likelihood is based on an assumed model and cannot account for bias sources that are not controlled by the model or the study design. Maximum likelihood is nonetheless popular, because it is computationally straightforward and intuitive and because maximum likelihood estimators have desirable large-sample properties in the (largely fictitious) case in which the model has been correctly specified. Here, we work through an example to illustrate the mechanics of maximum likelihood estimation and indicate how improvements can be made easily with commercial software. We then describe recent extensions and generalizations which are better suited to observational health research and which should arguably replace standard maximum likelihood as the default method.

epidemiologic methods; maximum likelihood; modeling; penalized estimation; regression; statistics

Abbreviations: CI, confidence interval; LASSO, least absolute shrinkage and selection operator; LR, likelihood ratio; ML, maximum likelihood; MSE, mean squared error.

Statistics is largely concerned with methods for deriving inferential quantities (such as estimates of unknown parameters) from observed data. Maximum likelihood (ML) may be the most widely used class of such methods in the health sciences. While brief descriptions of ML principles appear in some epidemiology textbooks (see *Modern Epidemiology* (1), chapter 13) and there are detailed descriptions in statistics textbooks, it is our experience that many epidemiologists have only a vague understanding of ML and other likelihood-based methods and their limitations. Here, we describe ML in enough detail to work through a simple example and discuss some key limitations of classical ML. We also describe extensions that can be used to address these limitations, including profile and penalized ML, and explain how the latter connects ML and Bayesian methods.

MODELS AND ML

For outcomes such as prevalence and incidence, a probability model is a formula that yields the probability of each observed value as a function of parameter values and measured covariates. Ideally, one should use a model that can reasonably approximate reality. By "reality" we mean the data distribution that would be produced by the mechanisms that generated the observed data (where the mechanisms include all sources of bias and departures from protocols, as well as intended design features). In practice, however, most analyses use models that are software defaults, such as logistic regression for proportions (risk or prevalence) and log-linear regression for incidence rates.

In most ML-based software, logistic model-fitting assumes binomial variation of the case count, while Poisson model-fitting assumes Poisson variation for the case count, leaving only the covariates to include in the model (and their form) to the investigator. Reality tends to be much more complicated than the resulting models, which usually depend on a rather small number of parameters. For purposes of illustration, however, we will maintain the fiction that the probability model is an adequate representation of reality.

The method of ML finds values of the model parameters, called ML estimates, which make the observed data most probable under the chosen model. In frequentist likelihood theory, the parameters are fixed constants that govern the distribution of the observations, and probabilities are viewed as hypothetical long-run frequencies (hence the term "frequentist"). We emphasize, however, that the ML estimates are usually not the most probable values of the parameters, because parameter probabilities are not part of likelihood theory. In Bayesian theory, parameters as well as data are treated as random variables, and thus Bayesian methods may be used to derive parameter probabilities (2, 3); however, our focus here will be frequentist methods.

AN EXAMPLE OF ML ANALYSIS

Consider a cohort study of incident diarrhea during a 10-day follow-up period among 30 infants, all colonized with *Vibrio cholerae* (see *Modern Epidemiology* (1), chapter 14). Our scientific objective is to estimate the 10-day risk of diarrhea among infants with low levels of antibiotics in maternal breast milk ($X = 1$) relative to infants with high levels ($X = 0$), as shown in Table 1. The cohorts are distinguished only by their values for X, and the data are the numbers falling ill, y_x, out of the total n_x, when $X = x$. Assuming that diarrhea occurs independently across cohort members and that the risk of diarrhea is p_x when $X = x$, the probability of seeing y_x cases in the cohort with $X = x$ is given by the binomial formula

$$\binom{n_x}{y_x} p_x^{y_x}(1-p_x)^{(n_x-y_x)},$$

where $\binom{n_x}{y_x}$ is the number of ways y_x cases can be chosen from n_x infants without regard to order (e.g., 6 choose $2 = \binom{6}{2} = 15$).

Now consider the logistic regression model $p_x = \text{expit}(\beta_0 + \beta_1 x)$, where $\text{expit}(u) = e^u/(1+e^u)$ is the logistic function. In this model, $e^{\beta_0} = p_0/(1-p_0)$ is the odds of being a diarrhea case in the unexposed group and e^{β_1} is $p_1/(1-p_1)/\{p_0/(1-p_0)\}$, the odds ratio associated with a 1-unit increase in exposure x. The likelihood at β_0, β_1 is defined as the probability that we would have seen the observed case numbers y_0, y_1 if the parameter values were truly β_0, β_1:

$$L(\beta_0, \beta_1; y_0, y_1) = \prod_{x=0,1} \binom{n_x}{y_x} p_x^{y_x}(1-p_x)^{(n_x-y_x)}.$$

If we denote the parameter list or vector (β_0, β_1) by $\boldsymbol{\beta}$ and the outcome list (y_0, y_1) by $\boldsymbol{y}$, we can compactly write this likelihood function as $L(\boldsymbol{\beta}; \boldsymbol{y})$. For example, at $\beta_0 = -1$, $\beta_1 = 1$, we get $\boldsymbol{\beta} = (-1,1)$, $p_0 = \text{expit}(-1) = e^{-1}/(1+e^{-1}) = 0.269$, $p_1 = \text{expit}(-1+1) = e^0/(1+e^0) = 0.500$, and $L(\boldsymbol{\beta}; \boldsymbol{y}) = \binom{16}{7} 0.269^7\, 0.731^9 \binom{14}{12} 0.5^{12} 0.5^2 = 0.000386$.

In finding $\boldsymbol{\beta}$ to maximize the likelihood, we can ignore factors like $\binom{n_x}{y_x}$ that do not depend on the unknown parameters. Additionally, the mathematics are easier if we instead work with the natural logarithm of the likelihood, which we will denote by $g(\boldsymbol{\beta})$:

$$g(\boldsymbol{\beta}) = \ln\{L(\boldsymbol{\beta}; \boldsymbol{y})\} = \sum_{x=0,1} y_x \ln(p_x) + (n_x - y_x)\ln(1-p_x).$$

The ML estimate of $\boldsymbol{\beta}$, denoted $\hat{\boldsymbol{\beta}}$, contains the values $\hat{\beta}_0$ and $\hat{\beta}_1$ for β_0 and β_1 that maximize (make as large as possible) $L(\boldsymbol{\beta}; \boldsymbol{y})$, or equivalently that maximize $g(\boldsymbol{\beta})$. One way to find $\hat{\boldsymbol{\beta}}$ is to differentiate $g(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. The resulting derivative is called the score function and is denoted $g'(\boldsymbol{\beta})$. The ML estimate $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)$ is a solution to $g'(\boldsymbol{\beta}) = 0$, which is called the score equation for $\boldsymbol{\beta}$. Because the first derivative of a function is its slope, we are seeking a point $\hat{\boldsymbol{\beta}}$ that makes the slope of $g(\boldsymbol{\beta})$ zero in all directions. This point can occur at a minimum of the function or at a maximum, as well as at other points; nonetheless, in typical regression analyses using generalized linear models (including linear, log-linear, logistic, and proportional hazards models), there will be only 1 solution to the score equation, $\hat{\boldsymbol{\beta}}$ (although some of the parameter estimates in $\hat{\boldsymbol{\beta}}$ may go to infinity), and the likelihood will be at a maximum at this solution.

Sometimes an ML estimate is a simple function of the data, known as a "closed-form solution." In our example, $\hat{\beta}_0$ is the sample log odds $\ln\{y_0/(n_0 - y_0)\} = \ln(7/9) = -0.251$ and $\hat{\beta}_1$ is the sample log odds ratio:

$$\hat{\beta}_1 = \ln\left\{\frac{y_1(n_0-y_0)}{y_0(n_1-y_1)}\right\} = \ln\left\{12 \times \frac{9}{(7\times 2)}\right\} = 2.043,$$

with standard error

$$\hat{s}_1 = \left\{\frac{1}{y_1} + \frac{1}{(n_0-y_0)} + \frac{1}{y_0} + \frac{1}{(n_1-y_1)}\right\}^{1/2} = 0.915$$

(1, p. 249). Nonetheless, most ML estimates must be found by means of iterative procedures. An appealing feature of ML is that estimates are unaffected by the choice of parameterization

Table 1. Distribution of 30 Infants According to Diarrhea Status During a 10-Day Follow-up of Breastfed Infants Colonized With *Vibrio cholerae*, Bangladesh, 1980–1981[a]

Antibiotic Level[b]	Diarrhea Cases ($y = 1$)	No Diarrhea ($y = 0$)	Total
Low ($x = 1$)	12	2	14
High ($x = 0$)	7	9	16
Total	19	11	30

[a] Cohort study of incident diarrhea during a 10-day follow-up among 30 infants, all colonized with *Vibrio cholerae* (source: *Modern Epidemiology* (1), chapter 14). The maximum likelihood estimate of the odds ratio ($\exp(\beta_1)$ in the logistic model) is $\{12 \times 9/(7 \times 2)\} = 7.71$.

[b] Marker for level of antibiotics in maternal breast milk.

(see *Theory of Point Estimation* (4), section 6.2), so that (for example) we would have obtained the same ML estimate for the log odds ratio had we taken the proportions exposed among cases and controls as our model parameters.

An important fact originally established for ML logistic regression is that all types of case-control data can be analyzed as if they were cohort data, using "multiplicative intercept" models. If there is no selection bias, only the resulting intercept estimate is distorted by the case-control sampling. For further discussion and citations, see *Modern Epidemiology* (1, pp. 429–435).

GENERAL ML THEORY

We now expand from the example to the general case. Suppose we are interested in studying the relationship of an outcome Y to a list (vector) of J covariates $\boldsymbol{X} = (X_1, \ldots, X_J)$ and have n observations with identification numbers (indices) $i = 1, 2, \ldots, n$. Let y_i and $\boldsymbol{x}_i$ be the value of Y and $\boldsymbol{X}$ for observation i. Parametric statistical modeling assumes that if observation i has covariate values x, then each y_i represents a random draw from a probability model $f(y|\boldsymbol{x}; \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is a list of unknown parameters; usually, $\boldsymbol{\beta}$ represents regression coefficients $(\beta_0, \beta_1, \ldots, \beta_J)$, where β_0 is the intercept and β_j is the coefficient of x_j. The model $f(y|\boldsymbol{x}; \boldsymbol{\beta})$ is a known function of $\boldsymbol{x}$ and $\boldsymbol{\beta}$ that returns the probability that $Y = y$ when $\boldsymbol{X} = \boldsymbol{x}$, given $\boldsymbol{\beta}$.

ML focuses directly on the probability of our observations as a function of the unknown parameters $\boldsymbol{\beta}$; that is, we study the properties of our model $f(y|\boldsymbol{x}; \boldsymbol{\beta})$ as a function of $\boldsymbol{\beta}$ rather than a function of Y, holding Y fixed at its observed (data) value. In other words, the value of the probability model for observation i, $f(y_i|\boldsymbol{x}_i; \boldsymbol{\beta})$, is treated as a function of the parameters $\boldsymbol{\beta}$ rather than as a function of the outcome Y or covariates X, and thus is written as $L(\boldsymbol{\beta}; y_i, x_i)$. This notation shift reminds us that we are looking at how $f(y|\boldsymbol{x}; \boldsymbol{\beta})$ varies with $\boldsymbol{\beta}$, instead of how it varies with Y. Below, we simplify $L(\boldsymbol{\beta}; y_i, x_i)$ to $L(\boldsymbol{\beta}; y_i)$ and thus leave the covariates implicit. $L(\boldsymbol{\beta}; y_i)$ is called the likelihood contribution from observation i.

When the individual outcomes y_i are independent of (not associated with) one another given the covariates (as is typically assumed for noncontagious diseases in unrelated individuals), the likelihood function is the product of the individual contributions:

$$L(\boldsymbol{\beta}; \boldsymbol{y}) = \prod_{i=1}^{n} L(\boldsymbol{\beta}; y_i) = \prod_{i=1}^{n} f(y_i|\boldsymbol{x}_i; \boldsymbol{\beta}),$$

where the boldface $\boldsymbol{y}$ represents all of the observed outcomes $(y_1, y_2, \ldots, y_n)$. We label $L(\boldsymbol{\beta}; \boldsymbol{y})$ a "likelihood function" rather than a probability model because it is not giving us probabilities for $\boldsymbol{\beta}$. Confusingly, the value of $L(\boldsymbol{\beta}; \boldsymbol{y})$ at a particular value $\boldsymbol{\beta}$ is sometimes called "the likelihood of $\boldsymbol{\beta}$ given the data," even though it is not a probability for $\boldsymbol{\beta}$.

If the model is correct and some technical conditions are satisfied (see Appendix 1), then the ML estimators are asymptotically unbiased and jointly normal, meaning that as the sample size n increases the distribution of $\hat{\boldsymbol{\beta}}$ can be approximated by a multivariate normal distribution with mean equal to the true value of $\boldsymbol{\beta}$. Epidemiologists may be more familiar with exact unbiasedness. The expected value (mean) of an exactly unbiased estimator equals the true parameter value regardless of n, but ML estimators of ratios of outcomes (such as the risk ratio and odds ratio) are only asymptotically unbiased.

Recall that 2-sided Wald confidence limits for a coefficient $\hat{\beta}_j$ are obtained by subtracting and adding a multiple (1.96 for 95% confidence) of its standard error $\hat{s}_j$. The ML estimator is as precise as any other asymptotically unbiased estimator that can be both constructed from the likelihood alone and used to center approximately valid Wald confidence intervals. Appendix 1 describes how standard errors for ML estimates can be derived by taking the second derivatives of the likelihood function to obtain what is known as the "information" matrix.

For those more familiar with least-squares regression, ML may seem very different. Nonetheless, in the case of normal linear regression, the least-squares estimate is identical to the ML estimate. In generalized linear modeling, ML regression can be viewed as an iterative refinement of weighted least squares in which the inverse-variance weights are updated and the model is refitted at each iteration using the parameter estimates from the previous iteration (5). These facts show that ML attempts to find the parameter values that bring the fitted model closest to the data points, in the same sense as does least squares (6).

PROFILE LIKELIHOOD

One useful approach to maximization is to set β_1 to a given value and then find the value of β_0 that maximizes the log-likelihood $g(\boldsymbol{\beta})$ given that value of β_1. We can repeat this computation across a range for β_1, $(\beta_1^{\min}, \beta_1^{\max})$, broad enough to include the maximum and in sufficiently small steps to have the desired resolution for β_1. For each value of β_1 in that grid, we obtain the maximum of $g(\boldsymbol{\beta})$ when β_1 is set to that step value. The resulting function $g(\beta_1)$ shows the maximum possible log-likelihood for each β_1 and is called a profile log-likelihood function for β_1; its antilog is the profile likelihood for β_1 and has its maximum at the ML estimate $\hat{\beta}_1$.

Profile likelihood is often used when accurate interval estimates are difficult to obtain using standard methods—for example, when the log-likelihood function is highly nonnormal in shape or when there is a large number of nuisance parameters (7). Usually there will be 2 values for β_1, β_1^{lower} and β_1^{upper}, where the profile likelihood is $e^{-3.84/2} = 14.7\%$ that of the ML estimate, where 3.84 is the 95th percentile of a 1-degree-of-freedom χ^2 variate. β_1^{lower} and β_1^{upper} are then approximate 95% confidence limits for β_1 and are called profile likelihood or likelihood ratio (LR) limits. When fitting a simple model to a large data set, Wald limits and LR limits will typically be similar. In more complex settings or with small sample sizes, LR can provide more accurate coverage because it does not depend on normality of the ML estimate $\hat{\beta}_1$ (8, p. 9). LR limits tend to be asymmetric in such settings. LR limits can be obtained through the SAS procedure GENMOD (SAS Institute Inc., Cary, North Carolina) by using the "LRCI" option in the model statement and in Stata using the "pllf" command

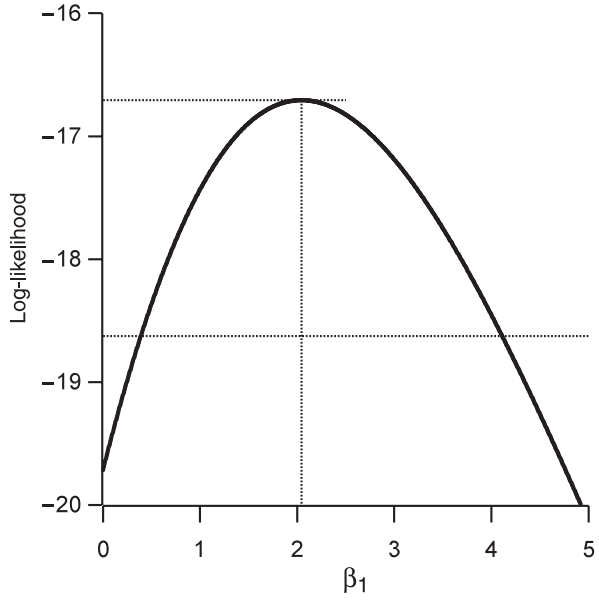

Figure 1. Profile log-likelihood for the log odds ratio, β_1.

(StataCorp LP, College Station, Texas) (9). Appendix 2 gives a method for plotting profile log-likelihoods.

Fitting the logistic model by ML to Table 1, we get $\hat{\beta}_1 = 2.043$ and $\hat{s} = 0.915$, which yields an odds ratio of 7.71. The Wald 95% confidence interval for the odds ratio is $\exp(2.043 \pm 1.96 \times 0.915) = (1.28, 46.3)$. The 95% LR confidence interval is 1.47, 61.1, as can be read off the profile log-likelihood plot in Figure 1 (or found numerically with the data used to generate the plot). The lower horizontal reference line is drawn at 3.84/2 = 1.92 units below the maximum log-likelihood. The ratio of upper to lower limits is 61.1/1.47 = 41.6 for the LR limits, 15% larger than the ratio 46.3/1.28 = 36.2 for the Wald limits, although in other examples the LR limits can be narrower.

Although the LR method is preferable to the Wald method, there is a third method for computing confidence limits and *P* values from the likelihood function, based on the score function $g'(\boldsymbol{\beta})$ and the expected information (defined in Appendix 1). This score method often has better small-sample properties than the Wald or LR method (10–12) and is described in Appendix 3, along with general LR methods.

PENALIZED LIKELIHOOD

Penalization is a method for circumventing problems in the stability of parameter estimates that arise when the likelihood is relatively flat, making determination of the ML estimate difficult by means of standard or profile approaches. Penalization is also known as shrinkage, semi-Bayes, or partial-Bayes estimation, although it does not require a Bayesian justification; instead, it can be viewed as a method for introducing some tolerable degree of bias in exchange for reduction in the variability of parameter estimates (13) (see *Modern Epidemiology* (1), chapter 21). Penalization can be applied to any estimation method, although here we focus on penalized likelihood and its extensions.

A penalized log-likelihood is just the log-likelihood with a penalty subtracted from it that will pull or shrink the final estimates away from the ML estimates, toward values $\boldsymbol{m} = (m_1, \ldots, m_J)$ that have some grounding in information outside of the likelihood as good guesses for the β_j in $\boldsymbol{\beta}$. The most common penalty is the sum of squared differences between the individual components of $\boldsymbol{\beta}$ and the individual components of $\boldsymbol{m}$, $\sum_{j=1}^{J}(\beta_j - m_j)^2$, known as a quadratic penalty and denoted here by $(\boldsymbol{\beta} - \boldsymbol{m})^2$. The penalized log-likelihood is then $\ln\{L(\boldsymbol{\beta}; \boldsymbol{y})\} - r(\boldsymbol{\beta} - \boldsymbol{m})^2/2$, where $r/2$ is the weight attached to the penalty relative to the original log-likelihood. We maximize this penalized log-likelihood to obtain the penalized ML estimate. From this formulation we can see that the penalty gets bigger rapidly as $\boldsymbol{\beta}$ gets further away from $\boldsymbol{m}$, and that the effect of the penalty on the final estimate (i.e., the difference between the ordinary ML estimate and the penalized ML estimate) is directly proportional to r (14).

In frequentist theory, a penalty function is a stabilization (smoothing) device to improve the repeated-sampling (frequency) performance of an estimator (14). The choice of penalty may be guided by background information—for example, that large values for the parameter are implausible. To understand the relationship of penalized likelihood to Bayesian theory, recall that a prior distribution for a given parameter is a probability distribution that describes our information about the parameter, outside of any information conveyed by the data under analysis (15); the term "prior" refers to this distribution or its density function. Priors with large variances represent limited or weak background information, while priors with small variances represent extensive background information. If we want our penalized estimates to accurately incorporate the background information in our prior, we take −2 times the log of the prior density as the penalty function. This practice is the source of the divisor of 2 in the above penalized-likelihood formula, and it makes r a measure of the prior information. Specifically, r is the precision (the inverse of the variance) of the parameters in $\boldsymbol{\beta}$ in the original prior distribution for those parameters. This interpretation assumes that the parameters in $\boldsymbol{\beta}$ have independent prior information with the same precision; this assumption can be relaxed using more general formulations as given below.

That both frequency and Bayesian theories can make use of priors shows that they can be viewed as complementary, not conflicting (13, 16). From a Bayesian perspective, quadratic log-likelihood penalization corresponds to having independent normal priors on the coefficients with prior means m_j and prior variance $1/r$; thus, cautious priors have small precision (large variance). From the frequentist perspective, if the m_j are well chosen, the same quadratic penalization is a method for reducing mean squared error (MSE), which is the average squared distance between the estimate and the correct value of $\boldsymbol{\beta}$, or $\text{MSE} = \text{bias}^2 + s^2$, where s is the standard deviation of the estimate.

Penalization will reduce MSE whenever it reduces s more than it increases bias. In addition, penalization will reduce

MSE whenever it reduces bias. In particular, penalization can reduce bias when the ML estimator is only asymptotically unbiased (as in logistic regression) and therefore subject to finite-sample bias (defined in the next section) (17, 18). The best value of r for penalization is unknown because it depends on how far the chosen value $\boldsymbol{m}$ is from the unknown correct value of $\boldsymbol{\beta}$. However, it will be between the extremes of using $\hat{\boldsymbol{\beta}}$ as our estimate (corresponding to $r = 0$ and thus ignoring the penalty function) and using $\boldsymbol{m}$ as our estimate (corresponding to an r value so large that the likelihood function and data are essentially ignored); hence, r is often called the "tuning" parameter, reflecting that the lowest MSE will ensue when r is carefully tuned between 0 and ∞. If there are many parameters and the data set is very large relative to the number of parameters, the best value of r or the variance $1/r$ can be estimated using cross-validation or empirical Bayes methods (14, 19); otherwise, as in our example, one may do better by examining results for values consonant with background information (20).

Suppose we think that $\boldsymbol{\beta} = (\beta_0, \beta_1)$ is not far from $\boldsymbol{m} = (\text{logit}(0.1), \ln(2))$, where $\text{logit}(p) = \ln\{p/(1-p)\}$. This information may come from the existing literature or consensus of colleagues. In our example, this penalty implies that we expect 10% of the infants with high levels of antibiotics to incur diarrhea and that having low levels of antibiotics doubles the odds of diarrhea. We must specify how much weight r to place on the penalty relative to the likelihood. Penalization will reduce the standard errors but will also add bias to the extent that $\boldsymbol{m}$ is incorrect, in proportion to r. If the penalty weight is large relative to the log-likelihood weight (r much larger than 1), we may incur unacceptable bias. Nonetheless, if we are cautious in our choice of prior location $\boldsymbol{m}$ and precision r, we expect penalization to reduce MSE.

There are many variations possible for penalty functions, including using a separate weight $\boldsymbol{r} = r_j$ for each coefficient $\boldsymbol{\beta} = \beta_j$, which amounts to specifying a separate prior variance $1/r_j$ for each β_j (21). More generally, r can be replaced by a weight matrix $\boldsymbol{R}$, which can be interpreted as a prior information matrix (inverse covariance) for $\boldsymbol{\beta}$ and which allows prior correlations among the β_j's (22); the penalized log-likelihood is then $\ln\{L(\boldsymbol{\beta}; \boldsymbol{y})\} - (\boldsymbol{\beta} - \boldsymbol{m})' \boldsymbol{R}(\boldsymbol{\beta} - \boldsymbol{m})/2$. Table 2 provides results from penalized estimation of the log odds ratio β_1 for Table 1, using $m_1 = 0$ and $r_1 = (0, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8, \infty)$, and gives the 95% prior intervals corresponding to each r_1. The 95% prior interval is an interval in which 95% of the prior probability distribution resides; in the present setting, this prior distribution is normal and centered at m_1. We place no penalty on the intercept ($r_0 = 0$), although one could easily do so. The resulting estimates of β_1 can be viewed as the result of shrinking the ML estimate $\hat{\beta}_1$ toward the prior mean log odds ratio m_1, or information-weighted averaging of $\hat{\beta}_1$ and m_1.

We used rescaled data augmentation (for details, see Greenland (2, 3) and Sullivan and Greenland (21)) as well as the SAS procedure NLMIXED (see Appendix 4) to obtain penalized estimates; as expected, these estimates were very close numerically. Various Bayesian and mixed-modeling software packages can be used instead, including the SAS procedure GLIMMIX (23). One may also use a profile penalized likelihood approach to obtain confidence intervals that have coverage closer to their stated 95% probability than the Wald intervals (21, 22).

Another popular penalty is the sum of absolute deviations $\sum_{j=1}^{J} |\beta_j - m_j| = |\boldsymbol{\beta} - \boldsymbol{m}|$, which corresponds to using double-exponential prior distributions (which, unlike the normal distribution, has heavy tails spread around a sharp peak at $\boldsymbol{m}$) and leads to least absolute shrinkage and selection operator (LASSO) regression based on the penalized log-likelihood $\ln\{L(\boldsymbol{\beta}; \boldsymbol{y})\} - r|\boldsymbol{\beta} - \boldsymbol{m}|$ (14). Unlike quadratic penalties, LASSO can shrink the β_j estimate all the way to m_j, thus deleting β_j from the model if $m_j = 0$ (this can happen because, unlike a quadratic penalty, the LASSO penalty can overwhelm the likelihood for ML estimates near m_j). Such methods have proven superior to ordinary ML logistic regression for

Table 2. Penalized Maximum Likelihood Estimates of the Odds Ratio[a] for the Example Cohort Data on Diarrhea During a 10-Day Follow-up of Breastfed Infants Colonized With *Vibrio cholerae*, Bangladesh, 1980–1981

r_1 (Precision)	95% Prior Interval for Odds Ratio	Penalized Likelihood			
		Data Augmentation	95% CI	NLMIXED	95% CI
0[b]	0, ∞	7.71	1.28, 46.4	7.71	1.28, 46.4
1/8	0.004, 256	6.40	1.22, 33.6	6.39	1.22, 33.5
1/4	0.020, 50.4	5.51	1.16, 26.2	5.51	1.16, 26.1
1/2	0.063, 16.0	4.39	1.07, 18.1	4.39	1.07, 18.1
1	0.141, 7.10	3.26	0.95, 11.2	3.25	0.95, 11.2
2	0.250, 4.00	2.34	0.83, 6.56	2.34	0.83, 6.56
4	0.375, 2.66	1.73	0.76, 3.94	1.73	0.76, 3.94
8	0.500, 2.00	1.38	0.74, 2.59	1.38	0.74, 2.59
$r_1 \to \infty$[c]	1, 1	1.00	1.00, 1.00	1.00	1.00, 1.00

Abbreviation: CI, confidence interval.
[a] Zero prior mean m_1 for the log odds ratio.
[b] Equal to unpenalized maximum likelihood estimate.
[c] Ignores data, gives back prior m_1.

association screening and data mining (14), for clinical prediction (24), and for fitting models with inverse-probability weights, especially when there are many covariates (25–28).

SPARSE-DATA BIAS, THE CURSE OF DIMENSIONALITY, AND SEMIPARAMETRIC MODELS

When the sample size is small or the data are sparse, ML may produce estimates that are skewed away from the true parameter values even if no other bias is present (29). This problem is called small- (or finite-) sample bias or sparse-data bias, depending on the situation, and there are real examples in which it is severe (30). Epidemiologists are aware of problems due to sparse-data bias in very small studies, but sparse-data bias appears to be less widely recognized when it occurs in larger studies. A common misconception is that this bias is addressed by matching and conditional logistic regression, but sparse-data bias can be severe with those methods even when there is a large number of matched sets with few model covariates (29).

The problems of correct model specification and sparse-data bias are intertwined. As we attempt to fit a more flexible and elaborate model to better adjust for large-sample biases (e.g., confounding), we invariably add more parameters and thus make the data sparse relative to the model. For example, large samples can become sparse if one uses ordinary ML to simultaneously adjust (or stratify) for many variables in an attempt to control confounding. As another example, large samples quickly become sparse in genome-wide association studies with hundreds of thousands of exposure variables (i.e., single-nucleotide polymorphisms). This problem, in which the number of model parameters grows too fast relative to the sample size, has been labeled the "curse of dimensionality" (31). Methods such as ordinary ML are ill-suited for high-dimensional problems because, to achieve approximate normality, they require large numbers at all exposure-outcome combinations relative to the number of model parameters.

If the number of model parameters is smaller than the sample size, there are various fixes for sparse-data bias, including Firth's correction (17), which is available in the SAS procedure LOGISTIC, as well as penalties based on the number of parameters (32, 33) and Bayesian methods using weak priors (18, 21, 29, 30). Another approach is to reduce the number of estimated parameters by using semiparametric models. These models are typically fitted through generalizations of ML. Best known is partial likelihood (34), which we now briefly introduce. Partial likelihood applies when the full likelihood depends on 2 distinct parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ and only $\boldsymbol{\beta}$ is of interest; $\boldsymbol{\lambda}$ is then called the nuisance parameter. If the full likelihood can be expressed as a product of 2 functions, the first involving only $\boldsymbol{\beta}$, and this function satisfies standard technical conditions, then statistical inferences about $\boldsymbol{\beta}$ can be obtained by treating the first function alone as if it were a full likelihood. Partial likelihood is called semiparametric rather than fully parametric because $\boldsymbol{\lambda}$ is not estimated and indeed may be arbitrarily complex, even infinite-dimensional. Estimators obtained by maximizing the partial likelihood retain the desirable asymptotic properties of ML estimators from the full likelihood, except possibly efficiency (34).

The classic example of partial likelihood arises in Cox proportional hazards modeling (35), where $\boldsymbol{\lambda}$ is an infinite-dimensional parameter representing the baseline hazard function. The advantages of focusing on $\boldsymbol{\beta}$ alone include not only dimension reduction but also potential computational ease; the disadvantage is that if $\boldsymbol{\beta}$ also appears in the second function along with $\boldsymbol{\lambda}$, information on $\boldsymbol{\beta}$ in that function is not used and the resulting inferences will not be fully efficient.

Estimating-equation methods (36) (which include ML as a special case, as well as least-squares, quasi-likelihood (5, 37), M-estimation (38), inverse-probability-weighted estimators (39), and targeted ML (40)) offer even more general approaches to semiparametric model-fitting. Such methods do not require specifying the entire probability distribution for the data, and they are overtaking ML as the standard fitting method in some settings. In particular, modern methods for fitting longitudinal treatment-effect models such as *g*-estimation and inverse-probability-of-treatment weighting (41) are based on estimating-equation theory.

Semiparametric estimates may still suffer from sparse-data bias or excessive variance (inefficiency) wherever they depend on normal approximations similar to those used for ML; in addition, they discard potentially relevant information that might be captured by using flexible parametric models (29, 42). To handle these difficulties, extensions of profile, score, and penalized methods to semiparametric modeling are available; for example, penalization may be used for the weight-estimation process in marginal structural modeling (25–28).

DISCUSSION

Perhaps the foremost reason ML is widely used in epidemiology is that it is the default method in commercial software for logistic, log-linear, and survival regression models. This software ubiquity may in turn be traced to the desirable computational and statistical properties of ML under the models ordinarily used in epidemiology. These properties include statistical consistency, asymptotic unbiasedness, asymptotic normality, and minimum variance among estimators with these large-sample properties. We note that even when the model is incorrect, a model parameter is still often interpretable as an approximate summary population parameter such as the logarithm of a population-averaged (marginal) or total-population standardized rate ratio (2–4).

If we let go of asymptotic unbiasedness, there are many estimation methods that are more accurate than ML, in the sense of retaining statistical consistency while having smaller MSE. This happens because the bias of competing estimators diminishes with sample size and is outweighed by reduction in standard error. In frequentist statistics, the bias introduced by penalization can be viewed as an adjustment for anticipated random error in the estimate; in multiparameter settings, the resulting accuracy improvement is called "Stein's paradox" (the paradox being that introducing bias improves accuracy) or the shrinkage effect (43).

The theoretical properties of any method do not ensure that the method performs well in analysis of epidemiologic studies. First, derivations of those properties assume that all or most aspects of the probability model are correct, yet such

perfect specification is essentially impossible when addressing real epidemiologic problems. A common response is that when a model is flexible enough to faithfully approximate the process under investigation and the sample size (given measured covariates) is adequate, the above properties should hold approximately. Nonetheless, in realistic-sized epidemiologic studies, being "flexible enough" can lead to sparse-data problems, which will call for penalization, semiparametric modeling, or some combination of methods beyond ML. Further flexibility to allow for uncontrolled bias can lead to an inability to estimate target parameters, which will necessitate the use of penalization or analogous Bayesian methods to obtain reasonable estimates (44).

ACKNOWLEDGMENTS

Author affiliations: Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina (Stephen R. Cole); Department of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota (Haitao Chu); Department of Epidemiology, Fielding School of Public Health, University of California, Los Angeles, Los Angeles, California (Sander Greenland); and Department of Statistics, College of Letters and Science, University of California, Los Angeles, Los Angeles, California (Sander Greenland).

Dr. Stephen Cole was supported in part by National Institutes of Health grant R01AI100654, and Dr. Haitao Chu was supported in part by National Institutes of Health grant R21AI103012.

Conflict of interest: none declared.

REFERENCES

1. Rothman KJ, Greenland S, Lash T. *Modern Epidemiology*. 3rd ed. New York, NY: Lippincott-Raven Publishers; 2008.
2. Greenland S. Bayesian perspectives for epidemiological research: I. Foundations and basic methods. *Int J Epidemiol*. 2006;35(3):765–775.
3. Greenland S. Bayesian perspectives for epidemiological research. II. Regression analysis. *Int J Epidemiol*. 2007;36(1):195–202.
4. Lehmann EL. *Theory of Point Estimation*. New York, NY: John Wiley & Sons, Inc; 1983.
5. McCullagh P, Nelder JA. *Generalized Linear Models*. London, United Kingdom: Chapman & Hall Ltd; 1989.
6. White H. *Estimation, Inference and Specification Analysis*. New York,NY: Cambridge University Press; 1994.
7. Greenland S. Likelihood-ratio testing as a diagnostic method for small-sample regressions. *Ann Epidemiol*. 1992;2(3):311–316.
8. Agresti A. *Categorical Data Analysis*. 2nd ed. New York, NY: John Wiley & Sons, Inc; 2002.
9. Royston P. Profile likelihood for estimation and confidence intervals. *Stata J*. 2007;7(3):376–387.
10. Agresti A. Score and pseudo-score confidence intervals for categorical data analysis. *Stat Biopharm Res*. 2011;3:163–172.
11. Vollset SE, Hirji KF, Afifi AA. Evaluation of exact and asymptotic interval estimators in logistic analysis of matched case-control studies. *Biometrics*. 1991;47(4):1311–1325.
12. Agresti A. On logit confidence intervals for the odds ratio with small samples. *Biometrics*. 1999;55(2):597–602.
13. Greenland S. Principles of multilevel modelling. *Int J Epidemiol*. 2000;29(1):158–167.
14. Hastie T, Tibsharani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York, NY: Springer Publishing Company; 2009.
15. Greenland S. Probability logic and probabilistic induction (with comment by Maclure). *Epidemiology*. 1998;9(3):322–332.
16. Gelman A, Carlin JB, Stern HS, et al. *Bayesian Data Analysis*. 2nd ed. Boca Raton, FL: CRC Press; 2003.
17. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika*. 1993;80(1):27–38.
18. Greenland S. Small-sample bias and corrections for conditional maximum-likelihood odds-ratio estimators. *Biostatistics*. 2000;1(1):113–122.
19. Efron B. *Large-scale Inference: Empirical Bayes Methods for Estimation, Testing and Prediction*. New York, NY: Cambridge University Press; 2010.
20. Greenland S. Methods for epidemiologic analyses of multiple exposures: a review and comparative study of maximum-likelihood, preliminary-testing, and empirical-Bayes regression. *Stat Med*. 1993;12(8):717–736.
21. Sullivan SG, Greenland S. Bayesian regression in SAS software. *Int J Epidemiol*. 2013;42(1):308–317.
22. Greenland S. Generalized conjugate priors for Bayesian analysis of risk and survival regressions. *Biometrics*. 2003;59(1):92–99.
23. Witte JS, Greenland S, Kim LL, et al. Multilevel modeling in epidemiology with GLIMMIX. *Epidemiology*. 2000;11(6):684–688.
24. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. New York, NY: Springer Publishing Company; 2008.
25. McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol Methods*. 2004;9(4):403–425.
26. Schneeweiss S, Rassen JA, Glynn RJ, et al. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*. 2009;20(4):512–522.
27. Westreich D, Lessler J, Funk MJ. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *J Clin Epidemiol*. 2010;63(8):826–833.
28. Lee BK, Lessler J, Stuart EA. Weight trimming and propensity score weighting. *PLoS One*. 2011;6(3):e18174.
29. Greenland S, Schwartzbaum JA, Finkle WD. Problems due to small samples and sparse data in conditional logistic regression analysis. *Am J Epidemiol*. 2000;151(5):531–539.
30. Hamra GB, MacLehose RF, Cole SR. Sensitivity analyses for sparse-data problems using weakly informative Bayesian priors. *Epidemiology*. 2013;24(2):233–239.
31. Robins JM, Ritov Y. Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Stat Med*. 1997;16(1–3):285–319.
32. Burnham KP, Anderson DR. *Model Selection and Multi-model Inference: A Practical Information-theoretic Approach*. 2nd ed. New York, NY: Springer-Verlag New York; 2002.
33. Claeskens G, Hjort NL. *Model Selection and Model Averaging*. New York, NY: Cambridge University Press; 2008.
34. Cox DR. Partial likelihood. *Biometrika*. 1975;62(2):269–276.
35. Cox DR. Regression models and life tables [with discussion]. *J R Stat Soc (B)*. 1972;34(2):187–220.
36. Godambe VD. *Estimating Functions*. New York, NY: Clarendon Press; 1991.

37. Wedderburn RWM. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*. 1974;61(3):439–447.
38. Stefanski LA, Boos DD. The calculus of M-estimation. *Am Stat*. 2002;56(1):29–38.
39. Tsiatis AA. *Semiparametric Theory and Missing Data*. New York, NY: Springer Publishing Company; 2006.
40. Van der Laan M, Rose S. *Targeted Learning: Causal Inference from Observational and Experimental Data*. New York, NY: Springer Publishing Company; 2011.
41. Robins JM, Hernán MA. Estimation of the causal effects of time-varying exposures. In: Fitzmaurice G, Davidian M, Verbeke G, et al, eds. *Longitudinal Data Analysis*. New York, NY: Chapman & Hall, Inc; 2008:553–597.
42. Whittemore AS, Keller JB. Survival estimation using splines. *Biometrics*. 1986;42(3):495–506.
43. Efron B, Morris C. Stein's paradox in statistics. *Sci Am*. 1977;236(5):119–127.
44. Greenland S. Relaxation penalties and priors for plausible modeling of nonidentified bias sources. *Stat Sci*. 2009;24(2):195–210.
45. Casella G, Berger RL. *Statistical Inference*. 2nd ed. Pacific Grove, CA: Duxbury Press; 2002.
46. Efron B, Hinkley DV. Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. *Biometrika*. 1978;65(3):457–487.
47. Maldonado G, Greenland S. A comparison of the performance of model-based confidence intervals when the correct model form is unknown: coverage of asymptotic means. *Epidemiology*. 1994;5(2):171–182.

APPENDIX 1

Wald Interval Estimation in Maximum Likelihood

For Wald intervals to be valid, the point estimator should be locally uniformly asymptotically unbiased normal, meaning that as the sample size grows, its distribution converges to a normal distribution with mean equal to the true value, with a discrepancy from normality that is proportional to $n^{-1/2}$ (where n is the sample size) and bounded within a neighborhood of the true value. Generally, a maximum likelihood estimator will satisfy this condition if the model $f(y_i|\mathbf{x}_i; \boldsymbol{\beta})$ satisfies certain technical or "regularity" conditions, which among other things ensure that (in probability) the likelihood, score, and information functions are finite smooth functions of the parameters and are not dominated by any single observation as the sample size grows (45, p. 516).

We also need variance estimates to obtain standard errors for interval estimation. The variances in the approximate sampling distribution for $\hat{\boldsymbol{\beta}}$ are inversely proportional to the sample size, which together with asymptotic unbiasedness implies that maximum likelihood estimators are also statistically consistent if the model is correct, meaning that the probability of $\hat{\boldsymbol{\beta}}$ falling within any given distance of the true value increases as n increases (called convergence in probability to the true value). The negative of the matrix of second derivatives of the log-likelihood $g(\boldsymbol{\beta})$, which is denoted $-g''(\boldsymbol{\beta})$ or $I(\boldsymbol{\beta})$, is called the observed information at $\boldsymbol{\beta}$. If the model is correct, the inverse of the observed information at $\hat{\boldsymbol{\beta}}$, $I(\hat{\boldsymbol{\beta}})^{-1}$, is an estimate of the covariance matrix of the approximate sampling distribution for $\hat{\boldsymbol{\beta}}$. The estimated standard error $\hat{s}_j$ for a coefficient β_j in $\hat{\boldsymbol{\beta}}$ is the square root of the corresponding diagonal of this estimated covariance matrix.

The expected information $E\{I(\boldsymbol{\beta})\}$ is often called the Fisher information at $\boldsymbol{\beta}$; its inverse evaluated at $\hat{\boldsymbol{\beta}}$, $E\{I(\hat{\boldsymbol{\beta}})\}^{-1}$, is also a covariance-matrix estimate for $\hat{\boldsymbol{\beta}}$ and is used in some generalized linear model software. In normal linear models, binomial logistic models, and Poisson log-linear models, the observed information and expected information are equal. Nonetheless, $I(\hat{\boldsymbol{\beta}})^{-1}$ may provide better approximate standard errors (46, 47) and a more intuitive derivation. From calculus, recall that the second derivative $g''(\hat{\boldsymbol{\beta}}) = -I(\hat{\boldsymbol{\beta}})$ measures the curvature of the likelihood function at $\hat{\boldsymbol{\beta}}$, with negative diagonal values representing downward curvature (indicating that $\hat{\boldsymbol{\beta}}$ is a maximum). A more curved log-likelihood function will be less spread out and thus will contain more information about the parameters. Note that changing the sign of $g''(\hat{\boldsymbol{\beta}})$ to obtain our information measure $I(\hat{\boldsymbol{\beta}})$ makes larger diagonal values in $I(\hat{\boldsymbol{\beta}})$ represent more curvature (a sharper likelihood peak), which in turn represents more information in the likelihood function and smaller variance estimates for $\hat{\boldsymbol{\beta}}$.

APPENDIX 2

Plotting Profile Likelihoods

One can find likelihood ratio limits by constructing multiple copies of the data set in which β_1 is set to different values between $\beta_1^{\min}$ and $\beta_1^{\max}$. For each data set, define the variable $\beta_1 x$ and declare this variable an offset (a variable in the model whose coefficient will not be estimated but instead will be set to 1); then run the regression analysis on this data set (but not otherwise including the exposure x). Next, plot the resulting profile log-likelihoods from each data set as a function of β_1, as shown in Figure 1 for our example. A SAS program to do this is as follows.

```
*Cohort data from ME3 Table 14.4;
data a; input x y n; cards;
1 12 14
0 7 16
;

*Profile ML;
data b;
  set a;
  do b1=0 to 5 by .005;
    b1x=b1*x;
    output;
  end;
proc sort data=b; by b1;
run; ods select none; run; *Turn off output
  from multiple logit models;
proc genmod data=b;
  by b1;
  model y/n=/d=b offset=b1x;
  ods output modelfit=c;
```

```
run; ods select all; run; *Turn output back
on;
data d;
  set c;
  by b1;
  logl=value;
  format logl stderr 10.4;
  if criterion="Log Likelihood" then out-
put;
  keep b1 logl;
*Next plot logl by b1;
```

APPENDIX 3

Likelihood-Ratio and Score Limits and Tests

Suppose now that we want to compare a simpler, more restricted model M_0 whose maximum likelihood is L_0 with a more complex reference model M whose maximum likelihood is L. We will consider the setting where the model M_0 is a special case of the model M with 1 or more parameters constrained, and so is "nested" within model M. Most often, M_0 is a version of M with 1 or more coefficients set to zero (i.e., 1 or more covariates dropped), but many other restrictions are possible; for example, M_0 may assume linearity while M allows more complex dose-response, in which the model M_0 assuming linearity is nested.

We can construct a P value comparing M_0 with M using the likelihood ratio (LR) statistic $-2\ln(L_0/L) = 2(\ln L - \ln L_0)$, which is also called the deviance of M_0 from M, denoted $\mathrm{Dev}(M_0, M)$. If M_0 is correct, then for sufficiently large samples this statistic is approximately χ^2 with degrees of freedom (df) equal to the difference in the number of unknown parameters in M and M_0. In the example, M is the original 2-parameter logistic model, while M_0 is the 1-parameter model without X, $\Pr(Y = 1|X = x) = \mathrm{expit}(\beta_0)$, which is restricted by $\beta_1 = 0$. From Figure 1 at $\beta_1 = \hat{\beta}_1$ and $\beta_1 = 0$, we see that $\ln(L) = -16.7067$, $\ln(L_0) = -19.7147$, and thus

$$\mathrm{Dev}(M_0, M) = 2\{-16.7067 - (-19.7147)\} = 6.016$$

on 1 df ($P = 0.014$). In comparison, the Wald statistic for $\beta_1 = 0$ is $2.043/0.915 = 2.23$ ($P = 0.026$). The 95% LR limits (1.47, 61.1) are the 2 values β_1^{lower} and β_1^{upper} for which the LR tests of the restrictions $\beta_1 = \beta_1^{\mathrm{lower}}$ and $\beta_1 = \beta_1^{\mathrm{upper}}$ give $P = 0.05$.

To describe score tests, suppose that the likelihood is maximized at $\tilde{\boldsymbol{\beta}}$ when we assume the restricted model M_0 (e.g., when $\beta_1 = 0$); then the expected information evaluated at $\tilde{\boldsymbol{\beta}}$, $\boldsymbol{E}\{\boldsymbol{I}(\tilde{\boldsymbol{\beta}})\}$, is an estimate of the covariance matrix for the score $g'(\tilde{\boldsymbol{\beta}})$ and can be used to construct an approximate χ^2 statistic for comparing M_0 with the unrestricted model M, with df equal to the difference in the number of unknown parameters in M and M_0. Pearson χ^2 statistics for 2-way tables, Mantel-Haenszel test statistics, and log-rank statistics for survival data are special cases. In our example, M_0 is the logistic model without X ($\beta_1 = 0$), and for comparing M_0 to the unrestricted model with X, the score is 3.133 and the expected information is 1.734; thus, the score statistic for $\beta_1 = 0$ is $3.133^2/1.734 = 5.662$ ($P = 0.017$), similar to the LR P value. Score 95% confidence limits for β_1 are the 2 values β_1^{lower} and β_1^{upper} for which the score tests of the restrictions $\beta_1 = \beta_1^{\mathrm{lower}}$ and $\beta_1 = \beta_1^{\mathrm{upper}}$ give $P = 0.05$.

APPENDIX 4

SAS Code for Penalized Maximum Likelihood in GENMOD and NLMIXED

```
*Cohort data from ME3 Table 14.4;
data a; input x y n; cards;
1 12 14
0 7 16
;

*PML by GENMOD using scaled data augmenta-
tion with r = 2;
data prior;
  s=10; int=0; h=0/s; x=1/s;
  r=2;
  v=1/r;
  y=(2/v)*s**2;
  n=2*y;
  output;
data a2; set a; h=0; int=1;
data a3; set a2 prior;
proc sort data=a3;
proc genmod data=a3;
  model y/n=int x/noint offset=h;
  ods select modelinfo modelfit parameter-
estimates;
  title "Penalized ML by scaled data aug-
  mentation, m=0 r=2";

*PML by NLMIXED using the same r;
*NOTE: # records is needed because SAS ap-
plies the penalty to each record;
*WARNING: Disregard generated CI because
it uses a t-multiplier based on an incorrect
df=2 (#records), instead create Wald CI using
the estimated SE and a standard normal mul-
tiplier (1.96 for 95% CI);
%macro pml(m,r,records);
  proc nlmixed data=a;
    parms b0 0 b1 0;
    p=1/(1+exp(-(b0+b1*x)));
    logl=(log(p)*y+log(1-p)*(n-y)) - .5*
    &r*(b1-&m)**2/&records;
    model y~general(logl);
    ods select specifications fitstatistics
parameterestimates;
    title "Penalized ML by NLMIXED, m=&m
r=&r";
%mend;
%pml(m=0,r=2,records=2);

run; quit; run;
```