Partial Likelihood
Author(s): D. R. Cox
Source: *Biometrika*, Vol. 62, No. 2 (Aug., 1975), pp. 269-276
Published by: Oxford University Press on behalf of Biometrika Trust
Stable URL: http://www.jstor.org/stable/2335362
Accessed: 03-05-2017 10:46 UTC

Biometrika (1975), **62**, 2, *p.* 269
Printed in Great Britain

Partial likelihood

BY D. R. COX
Department of Mathematics, Imperial College, London

SUMMARY

A definition is given of partial likelihood generalizing the ideas of conditional and marginal likelihood. Applications include life tables and inference in stochastic processes. It is shown that the usual large-sample properties of maximum likelihood estimates and tests apply when partial likelihood is used.

Some key words: Asymptotic theory; Censoring; Conditional likelihood; Life table; Marginal likelihood; Regression; Stochastic process.

1. INTRODUCTION

Likelihood is central to much theoretical discussion of statistical inference, from whatever viewpoint. In simple cases, the likelihood is just the joint density of the observed values considered as a function of the unknown parameters. The introduction of modified definitions for complicated problems is due to Bartlett (1937), notable recent contributions being by Fraser (1968), Kalbfleisch & Sprott (1970), and Andersen (1973).

Possible aims of modified likelihood functions include:

(i) the achievement of robustness;

(ii) the study of problems, especially in stochastic processes, for which the full likelihood is difficult or impossible to compute;

(iii) the need to develop methods based on second-moment properties rather than full distributional assumptions (Wedderburn, 1974);

(iv) the reduction of dimensionality in situations with many nuisance parameters.

This paper concentrates on (iv). The need for special discussion of this situation arises in the sampling theory and pure likelihood approaches especially because of the failure of the method of maximum likelihood as a general technique when there are many nuisance parameters. In a Bayesian approach it may be a convenient approximation and simplification to bypass the prior distribution of the nuisance parameters.

2. SOME DEFINITIONS

Consider a vector y of observations represented by a random variable Y having density $f_Y(y;\theta)$ and suppose that Y is transformed to new random variables (V, W) by a transformation not depending on the unknown parameter. We call $f_V(v;\theta)$ the marginal likelihood based on V and $f_{W|V}(w|v;\theta)$ the conditional likelihood based on W given $V = v$, both being considered as functions of θ; the usual notation for probability densities is used. These definitions are both special cases of the following definition of partial likelihood.

Let Y be transformed into the sequence

$$(X_1, S_1, X_2, S_2, \ldots, X_m, S_m), \tag{1}$$

where the components may themselves be vectors. The number m of pairs of terms may in some cases be random; alternatively we can imagine dummy variables added to complete the sequence up to the maximum conceivable m. The full likelihood of the sequence (1) is

$$\prod_{j=1}^{m} f_{X_j|X^{(j-1)},S^{(j-1)}}(x_j|x^{(j-1)},s^{(j-1)};\theta)\prod_{j=1}^{m} f_{S_j|X^{(j)},S^{(j-1)}}(s_j|x^{(j)},s^{(j-1)};\theta), \tag{2}$$

where $x^{(j)} = (x_1, \ldots, x_j)$ and $s^{(j)} = (s_1, \ldots, s_j)$. We call the second product the partial likelihood based on S in the sequence $\{X_j, S_j\}$.

The conditional likelihood based on W given $V = v$ corresponds to the special case

$$S_1 = W, \quad X_1 = V$$

and the marginal likelihood of V to the special case in which X_1 is a known constant, $S_1 = V$ and $X_2 = W$. Both marginal and conditional likelihoods are in a natural sense ordinary likelihoods for derived experiments, but the same is not true in general for partial likelihood; this is because of the way the conditioning events change. The relation between the definitions is considered further in §4.

The partial likelihood is useful especially when it is appreciably simpler than the full likelihood, for example when it involves only the parameters of interest and not nuisance parameters. The following problems now arise:

(*a*) to provide constructive procedures for finding useful partial likelihoods;

(*b*) to specify circumstances under which all or nearly all the relevant information is contained in the partial likelihood;

(*c*) to develop 'exact' small-sample significance tests and confidence limits from partial likelihoods;

(*d*) to develop large-sample theory based on a partial likelihood;

(*e*) to examine possible approximate versions of the factorization (2).

Of these only (*d*) and to some extent (*e*) will be considered in any detail in this paper.

One aspect of (*a*) which has been questioned by the referee is the uniqueness of the decomposition in any particular application. There are two aspects to this. First, it is conceivable that there are two entirely different factorizations leading to distinct but equally plausible partial likelihoods for the parameters of interest. If such a situation were to arise, one would in practice have to form some qualitative merging of the conclusions from the separate analyses. Secondly, it is clearly possible formally to move components from X into S and vice versa. The general principles to follow here are clear even if difficult to express mathematically:

(i) no omitted factor, i.e. one of the X's, should contain important information about the parameters of interest, so that the X's should all have distributions depending in an essential way on nuisance parameters;

(ii) incidental parameters, and so far as possible nuisance parameters, should not occur in the partial likelihood.

At least in the examples, (i) and (ii) fix the appropriate factorization.

3. Some examples

The following examples all have an element of time ordering which motivates the factorization (2).

Example 1. Consider a discrete time stochastic process with discrete states $\{0, 1, \ldots\}$, the state 0 having special properties. Suppose that conditionally on state 0 being occupied at time n, and independently of the previous history of the process, the probability that it will be occupied at time $n+1$ is $p_n(\theta)$, where θ is an unknown parameter. Suppose further that on leaving state 0 the system executes 'tours' determined by a probability mechanism of complicated and possibly unknown form. If large-sample arguments are to be used, we suppose that the number of distinct sojourns in state 0 is large.

The decomposition (2) is as follows:

X_1 gives the path up to and including the first entry to state 0;

S_1 specifies the number of subsequent periods up to the next exit from state 0 but does not specify the state to which the system then moves;

X_2 gives the state entered on leaving 0 and the subsequent path up to and including the second entry to state 0,

and so on. The partial likelihood based on S in the sequence $\{X_j, S_j\}$ is then

$$\Pi\left[\{1-p_{e_j+r_j}(\theta)\}\prod_{k=0}^{r_j-1} p_{e_j+k}(\theta)\right],$$

where the jth sojourn in state 0 starts at entry time e_j and ends at time e_j+r_j. If the probability of remaining in state 0 is always θ, this is $\Pi_j\{(1-\theta)\,\theta^{r_j-1}\}$ and is the same as the marginal likelihood of the sojourns in state 0.

This example can be generalized in various ways. For instance, in a semi-Markov process it may be convenient to take X_j to be the time between the $(j-1)$th and jth transitions and S_j to be state occupied after the jth transition.

Example 2. Consider a grouped life table. There are initially n individuals at risk of failure; individuals may also be censored, or withdrawn, from the study. Thus in the jth cell the individuals at risk may fail, may be censored or may survive to the start of the $(j+1)$th cell. It is possible to calculate likelihood functions by following the behaviour of individuals; the alternative followed here is to work with the frequencies of the various categories of response.

Let n_j be the observed number at risk in the jth cell, the observed numbers of individuals failing and being censored being respectively s_j and q_j; thus $n_{j+1} = n_j - s_j - q_j$ with $n_1 = n$, the total number of individuals. To simplify the argument we make initially the normally unrealistic assumption that censoring takes place instantaneously at the end of the interval, so that the number of individuals at risk of failure is n_j throughout the interval. Denote by ϕ_j the probability that an individual at risk in the jth interval fails in that interval. Failures of different individuals are assumed independent.

The mechanism of censoring is not specified. The only requirement is that the probabilities ϕ_j are relevant to the whole population of individuals under study; thus if an individual censored in an earlier interval had survived to the start of the jth interval, his probability of failure would have been ϕ_j. This is a strong assumption to the effect that censoring and failure are determined by independent mechanisms. It does, however, encompass a wide variety of censoring mechanisms in which the number censored at the end of the jth interval depends in an arbitrary way on the previous numbers failed and censored. Very special cases are that in which each individual has a preassigned censoring time which operates if the individual has not failed by that time and that in which censoring time is a random variable distributed independently of failure time.

Let the random variable X_j represent the number of individuals censored just before the end of the $(j-1)$th interval and S_j represent the number of failures in the jth interval. Then the partial likelihood based on S in the sequence $\{X_j, S_j\}$ is

$$\Pi \binom{n_j}{s_j} \phi_j^{s_j}(1-\phi_j)^{n_j-s_j}. \tag{3}$$

In this the parameter of interest may be the sequence $\{\phi_1, \phi_2, \ldots\}$; alternatively these may be expressed in terms of a smaller number of unknowns.

Example 3. One possible model for the analysis of failure data when explanatory variables are available is to suppose that the hazard function for an individual at risk at age t and having a vector $z^{\mathrm{T}} = (z_1, \ldots, z_p)$ of explanatory variables is

$$\exp(\beta^{\mathrm{T}} z)\, \lambda_0(t), \tag{4}$$

where $\beta^{\mathrm{T}} = (\beta_1, \ldots, \beta_p)$ is a vector of unknown parameters and $\lambda_0(t)$ is an unknown arbitrary nonnegative function of time giving the hazard when $z = 0$. The explanatory variables are possibly time dependent. Individuals are subject to censoring under similar assumptions to Example 2.

We deal here only with continuous time and assume that failures occur at distinct times $t_{(1)} < \ldots < t_{(m)}$. Let $\mathscr{R}_j$ denote the risk set at time $t_j - 0$, that is the set of individuals who have not failed or been censored by that time. Further let z_k denote the value of z for the kth individual and $z_{(j)}$ the value for the individual failing at time $t_{(j)}$. Then Cox (1972) gave

$$\prod_j \frac{\exp(\beta^{\mathrm{T}} z_{(j)})}{\sum\limits_{k \in R_j} \exp(\beta^{\mathrm{T}} z_k)} \tag{5}$$

as a likelihood for inference about β, rather misleadingly calling it a conditional likelihood. Kalbfleisch & Prentice (1973) showed that (5) is a marginal likelihood of ranks under the very restrictive assumptions, that there is no censoring, and that z does not depend on time; see also Crowley (1974).

In fact (5) is a partial likelihood in the sense of §2. We take X_j to specify the censoring in $[t_{(j-1)}, t_{(j)})$ plus the information that a failure occurs for the first time at $t_{(j)}$; S_j specifies the particular individual that fails at $t_{(j)}$.

In all these examples the desire to consider the partial likelihood stems from the complexity of the leading factor in the full likelihood (2). This factor depends on nuisance parameters so irrevocably that it is hard to see how useful information about the parameter of interest can be extracted from it. Of course the position would change if, for instance, in Example 3, $\lambda_0(t)$ were severely retricted by a parametric assumption. Then a contribution from the probability of zero failures should be included from the intervals containing no failures.

Example 4. Suppose that in the model (4), $\lambda_0(t)$ is assumed constant; that is, an individual with constant z has exponentially distributed failure time. We now take a sequence $\{X_j, S_j\}$

in which X_j specifies a censoring and S_j the failure experience between censorings. Then, under the weak assumptions about censoring of Example 2, the partial likelihood is

$$\Pi \exp\left(-\lambda \int_0^{t_k} e^{\beta^{\mathrm{T}} z_k} dt\right) \Pi' \lambda e^{\beta^{\mathrm{T}} z_{(j)}}. \qquad (6)$$

The first product is over all individuals and t_k is the time of failure or censoring for the kth individual; z_k is in general a function of time. The second product is over failures and $z_{(j)}$ is evaluated at the instant of failure.

4. Relation with marginal and conditional likelihood

Except in very special cases the partial likelihood based on S in the sequence $\{X_j, S_j\}$ is not the same as the marginal likelihood based on S or as the conditional likelihood based on S given $X = x$. To see this, note first that the marginal likelihood based on S is

$$\Pi f_{S_j|S^{(j-1)}}(s_j|s^{(j-1)}; \theta),$$

and for this to be identical to the partial likelihood the sequences $\{X_j\}$ and $\{S_j\}$ must be independent.

Similarly the conditional likelihood based on S given $X = x$ is

$$\Pi f_{S_j|S^{(j-1)}, X}(s_j|s^{(j-1)}, x; \theta),$$

and this is identical with the partial likelihood if and only if for all j given

$$S^{(j-1)} = s^{(j-1)}, \quad X^{(j)} = x^{(j)},$$

then S_j is independent of $(x_{j+1}, \ldots)$.

These conditions are not satisfied in general in the examples. Note, however, that the conditions above are for exact equality of densities; proportionality as functions of θ is a weaker requirement which has not been investigated.

5. Large-sample theory

We now consider in outline maximum likelihood estimates and tests based on partial likelihood. While formal asymptotic theorems will not be proved, it will appear that under very broad conditions the usual properties hold. Central to the discussion is the efficient score from the partial likelihood, namely $U_. = \Sigma U_j$, where

$$U_j = U_j(c_j; \theta) = \frac{\partial \log f_{S_j|C_j}(S_j|c_j; \theta)}{\partial \theta};$$

the conditioning variables for S_j have been written C_j. Purely for notational simplicity we suppose θ to be one dimensional.

Provided that the usual regularity conditions hold for the conditional density of S_j given $C_j = c_j$, we have by differentiation under the expectation sign that

$$E(U_j|C_j = c_j) = 0; \qquad (7)$$

therefore unconditionally, whatever the distribution of C_j,

$$E(U_j) = 0.$$

Further if $j < k$, the condition $C_k = c_k$ implies that U_j is fixed. Hence

$$E(U_k|U_j = u_j) = 0 \quad (j < k), \tag{8}$$

so that unconditionally

$$E(U_j U_k) = 0 \quad (j \neq k).$$

Finally, again working with the conditional distribution of S_j given $C_j = c_j$, we have under the usual regularity conditions for this distribution that

$$\operatorname{var}(U_j|C_j = c_j) = E\left\{-\frac{\partial^2 \log f_{S_j|C_j}(S_j|c_j;\theta)}{\partial\theta^2}\middle| C_j = c_j\right\} = i_j(c_j), \tag{9}$$

say. Thus the unconditional variance is, by (7) and (9),

$$\operatorname{var}(U_j) = E\{i_j(C_j)\} = i_j,$$

say. Therefore

$$E(U_.) = 0, \quad \operatorname{var}(U_.) = \Sigma i_j. \tag{10}$$

Write $\mathscr{I}(\theta)$ for the observed value of minus the second derivative of the log partial likelihood. Then under mild conditions implying some degree of independence between the different U_j's and that the information values are not too disparate, we have, for large m, that

(*a*) $U_.$ is asymptotically normal with zero mean and variance Σi_j;

(*b*) $\mathscr{I}(\theta)/m$ and $\mathscr{I}(\hat{\theta})/m$ converge in probability to $\Sigma i_j/m$.

Thus $\{\mathscr{I}(\hat{\theta})\}^{\frac{1}{2}} U_.$ has asymptotically the standard normal distribution.

Finally, under weak conditions on the third derivative of the log likelihood, this implies that

$$(\hat{\theta}-\theta)\mathscr{I}^{\frac{1}{2}}(\hat{\theta}) \tag{11}$$

has the standard normal distribution, with an immediate generalization to the multiparameter case. By the same arguments, tests based on the maximum likelihood ratio method and an asymptotic chi-squared distribution are justified.

The explicit calculation of the i_j and the development of exact tests requires specification of the distribution of the conditioning variables. It is, however, central to the usefulness of partial likelihood that such specification is unnecessary in using the large-sample results.

The above asymptotic arguments hold as $m \to \infty$. Alternatively there may be another quantity n such that as $n \to \infty$ each of the efficient score components is asymptotically normal, the result (11) then following as $n \to \infty$ for fixed m. An instance of this is provided by the simple life table, Example 2, in which the number of cells is fixed but the total number of individuals at risk, n, is large. It then follows that the estimates $\hat{\phi}_j = s_j/n_j$ have asymptotic covariance matrix estimated by $\operatorname{diag}\{\hat{\phi}_j(1-\hat{\phi}_j)/n_j\}$; the asymptotic independence (Kaplan & Meier, 1958) follows because the mixed second derivatives of the log partial likelihood are identically zero.

In many ways the crucial step in the above argument is that leading to (8); the final results would apply to other derived likelihoods formed from the product of many factors, provided that the relation between the conditioning variables in the different factors is such as to ensure (8).

6. Approximate forms

In some cases it may be possible to extend the applicability of the above ideas by replacing the partial likelihood by a suitable approximation. Only one example will be given, connected with Example 2. In § 3 it was assumed in deriving (3) that censoring takes place instantaneously at the end of each interval.

If the number of individuals censored in a particular interval is appreciable, it will be desirable if possible to obtain fuller information about the precise points of censoring; otherwise, in the total absence of further information, it will be sensible to form upper and lower estimates of the relevant ϕ assuming that censoring is concentrated respectively at the beginning and at the end of the interval. Quite often, however, it will be reasonable to suppose both that the proportions of failures and censorings in any one cell are small and that events of both types occur through the interval in independent Poisson processes.

Consider then one interval with n individuals at risk; without loss of generality take the interval to be of unit length. Suppose that failure and censoring occur in independent Poisson processes of rates ρ_ϕ and ρ_λ respectively. Then $\phi = -\log(1-\rho_\phi)$. We observe the trinomial variable giving the numbers s, q and $n-s-q$ of failures, censored and surviving individuals. Unfortunately 'exact' inference about ϕ with ρ_λ as a nuisance parameter does not seem possible, but we may consider the conditional distribution of the number of failures given the number censored; this is binomial with index $n-q$ and parameter

$$\frac{\rho_\phi(1-e^{-\rho_\phi-\rho_\lambda})}{\rho_\phi+\rho_\lambda e^{-\rho_\phi-\rho_\lambda}} = \phi(1+\tfrac{1}{2}\lambda),$$

where $\lambda = -\log(1-\rho_\lambda)$ is the probability of censoring in the absence of failure, and third-order terms are ignored. Now λ is estimated approximately by q/n so that in the modified theory we replace (3) by

$$\Pi \binom{n_j-q_j}{s_j} \{\phi_j(1+\tfrac{1}{2}q_j/n_j)\}^{s_j} \{1-\phi_j(1+\tfrac{1}{2}q_j/n_j)\}^{n_j-q_j-s_j}. \tag{12}$$

This is essentially equivalent to the common device of treating the number at risk of failure as $n_j - \tfrac{1}{2}q_j$.

This modification is sensible when the modification from ϕ_j to $\phi_j(1+\tfrac{1}{2}q_j/n_j)$ makes a nontrivial difference and when the error arising from replacing λ_j by q_j/n_j is small compared with the random errors of estimation of ϕ_j. A formal treatment proceeds via an entirely notional sequence of problems in which as $n \to \infty$ the number of cells increases so that ϕ_j and λ_j tend to zero. In fact if the number of cells $\sim n^a$, then ϕ_j and $\lambda_j \sim n^{-a}$, where $\sim$ here denotes the asymptotic order of magnitude; thus $\operatorname{var}(\hat{\phi}_j) \sim n^{-a-1}$, the bias correction to $\phi_j \sim n^{-2a}$ and the error in the bias correction $\sim n^{-\frac{3}{2}a-\frac{1}{2}}$. Thus the bias correction as a multiple of the standard error $\sim n^{-2a}/n^{-\frac{1}{2}a-\frac{1}{2}} = n^{-\frac{3}{2}a+\frac{1}{2}}$ and the error in the bias correction from estimating λ_j, again as a multiple of the standard error, $\sim n^{-\frac{3}{2}a-\frac{1}{2}}/n^{-\frac{1}{2}a-\frac{1}{2}} = n^{-a}$. Thus, if for example $a = \frac{1}{4}$, the bias correction is large compared with the standard error and the error in the bias correction negligible.

References

Andersen, E. B. (1973). *Conditional Inference and Models for Measuring.* Copenhagen: Mentalhygienjnisk Forlag.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proc. R. Soc.* A **160**, 268–82.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *J. R. Statist. Soc.* B **34**, 187–220.

Crowley, J. (1974). A note on some recent likelihoods leading to the log rank test. *Biometrika* **61**, 533–8.

Fraser, D. A. S. (1968). *The Structure of Inference.* New York: Wiley.

Kalbfleisch, J. D. & Prentice, R. L. (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika* **60**, 267–78.

Kalbfleisch, J. D. & Sprott, D. A. (1970). Application of likelihood methods to models involving large numbers of parameters (with discussion). *J. R. Statist. Soc.* B **32**, 175–208.

Kaplan, E. L. & Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Am. Statist. Assoc.* **53**, 457–81.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models and the Gauss–Newton method. *Biometrika* **61**, 439–47.

[*Received and revised January* 1975]