ON PARTIAL LIKELIHOOD

N. REID

Abstract

Partial likelihood, introduced in Cox (1975), formalizes the construction of the inference function developed in Cox (1972) and referred there to as a conditional likelihood. Partial likelihood can also be viewed as a version of composite likelihood, a different example of which was studied in Cox and Reid (2004). In this note I describe the links between partial and composite likelihood, and the connections to profile, marginal, and conditional likelihood. Somewhat tangentially, two recent applications of the Cox proportional hazards model from the medical literature are briefly discussed, as they highlight the model's ongoing relevance while also raising some more general questions about inference.

Keywords: conditional likelihood; logistic regression; marginal likelihood; nuisance parameters; profile likelihood; posterior distributions

1. Introduction

Although I wasn't fortunate enough to be in the audience for the presentation of Cox (1972), even as a student I was aware of an unusual level of excitement among my professors over a new breakthrough "by D.R. Cox". As an MSc student in 1974, I was tasked with analysing some survival data, and my supervisor, Jim Zidek, managed to procure a stack of computer cards that encoded a Fortran program written by Norman Breslow to fit the proportional hazards model. I was to apply this to a set of censored time-to-failure data that had been collected by a surgeon pioneering new methods for coronary artery bypass grafts: I think this speaks to the speed with which this new model and the ideas behind it were taken up.

The development of partial likelihood (Cox, 1975) clarified some theoretical results of the 1972 paper. The abstract states "A definition is given of partial likelihood, generalizing the ideas of conditional and marginal likelihood." In Cox (1972) the function proposed for inference about the regression parameters was described as a conditional likelihood. Questions were raised in the discussion, particularly by Kalbfleisch and Prentice (1972), and by Breslow (1972), about the correctness of this description. When I later asked David about this, he grumbled a little bit, saying, "it really was *a* conditional likelihood; it was a form of conditional likelihood." (Reid, 1994).

Corresponding author: Department of Statistical Sciences, University of Toronto, 700 University Ave, 9th Floor, Toronto, Ontario M5G 1Z5, Canada nancym.reid@utoronto.ca.

In the following I describe the role of conditional and marginal likelihood functions in models with many nuisance parameters, before describing partial likelihood and its connection to these functions. Briefly, I describe a more general construction, pseudo- or composite likelihood, versions of which have been proposed for inference in models with complex dependence, and then close with brief discussion of the use of the Cox model in two recent medical applications.

2. Nuisance parameters

In general, in regression-style parametric inference, we have a model for a sample of observations, usually expressed in the form of a conditional density $f(y \mid X; \theta)$, where y is an $n \times 1$ vector and X is an $n \times p$ matrix of potentially explanatory variables. The parameter θ can often be usefully separated into parameters of immediate interest, ψ, and nuisance parameters, λ, included to make the model more appropriate for the application at hand. Assuming the n responses are independent, the joint density is

$$f(y \mid X; \theta) = \prod_{i=1}^{n} f(y_i \mid x_i; \theta), \tag{1}$$

where x_i is the ith row of X. The log-likelihood function for $\theta = (\psi, \lambda)$ is

$$\ell(\psi, \lambda; y) = \sum_{i=1}^{n} \log f(y_i \mid x_i; \psi, \lambda). \tag{2}$$

The appearance of a sum of independent components in (2) enables application of a central limit theorem in regular models, which gives approximate inference based on derived quantities. In the expressions below we will often omit in the notation the dependence on X, as the usual regression analysis is conditional on X.

The profile log-likelihood function is obtained by maximization of (2) with respect to the nuisance parameter λ, for ψ fixed:

$$\ell_{\mathrm{p}}(\psi) = \ell(\psi, \hat{\lambda}_{\psi}); \tag{3}$$

here the dependence of the profile log-likelihood function on the data is implicit. Two approximations useful for inference are the normal approximation to the distribution of the maximum likelihood estimator $\hat{\psi}$ and the χ^2 approximation to the distribution of the log-likelihood ratio statistic:

$$\hat{\psi} \dot{\sim} N\{\psi, j_{\mathrm{p}}^{-1}(\hat{\psi})\}, \quad 2\{\ell_{\mathrm{p}}(\hat{\psi}) - \ell_{\mathrm{p}}(\psi)\} \dot{\sim} \chi^2_{\nu}, \tag{4}$$

where $j_{\mathrm{p}}(\psi) = -\ell''_{\mathrm{p}}(\psi)$ is the negative Hessian of the profile log-likelihood function, and ν, the degrees of freedom for the χ^2 approximation, is the dimension of the parameter of interest ψ. These approximations apply when sampling from the model $f(y; \psi, \lambda)$, and under regularity conditions on this model.

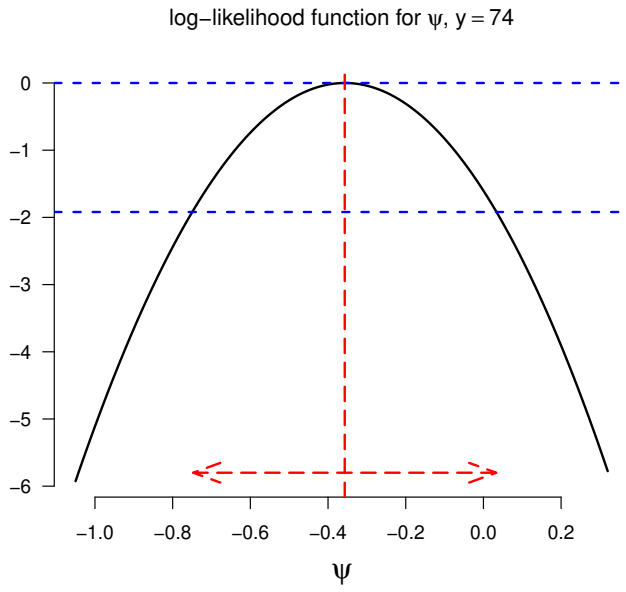

FIGURE 1. The profile log-likelihood function for the odds ratio in the 2×2 table discussed in §5. Deviations of ψ from the maximum likelihood estimate $\hat{\psi}$ can be measured on the normal scale (red; long dashed) or the χ^2 scale (blue; dashed).

In Figure 1 the profile log-likelihood function is plotted for a particular set of data from a 2×2 table, and illustrates how the approximations in (4) reflect deviations of the parameter of interest from the maximum likelihood value on two different scales. In this figure ψ is the log of the odds ratio; see §5. In Figure 2 we show some output from fitting a logistic regression model in `R` (R Core Team, 2021). The table of coefficient estimates and their approximate standard errors is based on the first approximation in (4). The construction of confidence intervals via profiling obtained using the `confint` function relies on the second approximation in (4). In both approximations each regression coefficient is treated in turn as the scalar parameter of interest.

Profile likelihood inference based on (4) can be unsatisfactory when the number of nuisance parameters is large; essentially the profile log-likelihood function is too concentrated around the maximum point, because errors in estimation of λ are not accounted for. A familiar example of this arises in inference about the variance parameter in a normal-theory linear regression. The maximum likelihood estimate of σ^2 is RSS/n, where RSS is the residual sum of squares, but the much-preferred, and widely-used, estimate is $RSS/(n-p)$, where p is the number of parameters in the linear regression.

This problem with profile likelihood inference was recognized by Fisher (1935), Neyman and Scott (1948), and many others. Improvements on the theory have been suggested using conditional or marginal arguments. In a general formulation, we search for a factorization of the joint density of y that isolates the parameter

```
> summary(myglm)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.079      0.987   -3.12   0.0018 **
aged1         -0.292      0.754   -0.39   0.6988
stage1         1.373      0.784    1.75   0.0799 .
grade1         0.872      0.816    1.07   0.2850
xray1          1.801      0.810    2.22   0.0263 *
acid1          1.684      0.791    2.13   0.0334 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be
1)

    Null deviance: 40.710  on 22  degrees of freedom
Residual deviance: 18.069  on 17  degrees of freedom

> confint(myglm)
Waiting for profiling to be done...
              2.5 % 97.5 %
(Intercept) -5.3002 -1.362
aged1       -1.7957  1.215
stage1      -0.1313  3.000
grade1      -0.7347  2.539
xray1        0.2669  3.523
acid1        0.2090  3.379
```

FIGURE 2. Some standard output illustrating a logistic regression model, with several explanatory variables. The z-value associated with each coefficient is based on the first approximation in (4); the confidence interval relies on the second approximation in (4), and in this instance the residual deviance provides a global test of the binomial model using the extension of (4) to testing several parameters of interest. Data taken from Davison (2003, Table 10.8).

of interest: i.e. we try to find statistics t_1 and t_2 so that either

$$f(y;\psi,\lambda) \propto f_m(t_1;\psi,\lambda) f_c(t_2 \mid t_1;\psi), \tag{5}$$

or

$$f(y;\psi,\lambda) \propto f_m(t_1;\psi) f_c(t_2 \mid t_1;\psi,\lambda). \tag{6}$$

The conditional likelihood function $L_c(\psi) \propto f_c(t_2 \mid t_1;\psi)$ in (5) or the marginal likelihood function $L_m(\psi) \propto f_m(t_1;\psi)$ in (6) can be used for inference about the parameter of interest using the same form of approximation as in (4), now based on the conditional (marginal) maximum likelihood estimate, or the conditional (marginal) log-likelihood ratio statistic. In special classes of models these factorizations arise from study of their structure. For example, factorization (5) can be established directly in linear exponential family models. Battey et al. (2023) considers techniques for deriving marginal versions (6) directly from the original model density $f(y;\psi,\lambda)$.

Conditional inference for the log-odds ratio in a 2×2 table conditions on the row and/or column totals. Fisher's exact test uses the hypergeometric distribution for one entry of the table, instead of the approximations in (4). Battey (2024) describes the conditional analysis for the more general case of logistic regression developed in Cox (1958b). An example of marginal likelihood is the `REML` version of generalized linear models with variance components. In the linear regression model mentioned above this leads to the usual unbiased estimate of σ^2.

Implicit in the use of conditional or marginal likelihood functions for inference is the assumption that the omitted density does not carry any (or very much) information about the parameter of interest, when the nuisance parameter is unknown. Making this precise has turned out to be difficult. Fisher (1935) wrote "if it be admitted that these marginal frequencies supply no information ... as to the proportionality of the frequencies in the body of the table" and referred to the marginal frequencies as 'ancillary', although it is in fact more correct to say that the cell entry is 'conditionally sufficient' for the nuisance parameter, as in this model factorization (5) applies.

In a model for survival time responses y_i, possibly censored, the parameters of interest are the regression parameters β. The failure and censoring processes, operating in continuous time, are the nuisance parameters. Cox's (1972) construction of the inference function for β was based on a conditional argument, but not on a conditional likelihood function, i.e. a conditional density for an identifiable function of the data. At each successive failure time the conditional probability that the observed individual failed, given the risk set of individuals available to fail, can be shown to be free of the unknown baseline hazard rate. The product of these conditional probabilities forms the function to be used for inference; as the risk sets at each failure time overlap, this does not correspond to a single conditional distribution, the point raised in the discussion. In Cox (1975) the function was called instead a partial likelihood.

3. Partial likelihood

Following Cox (1975), we might very generally think of two sequences of observations evolving in time, and observed at $j = 1, \ldots, n$ time points:

$$(X_1, S_1, X_2, S_2, \ldots, X_j, S_j, \ldots, X_n, S_n). \tag{7}$$

The joint density of the sequence can be expressed as

$$f(x, s; \psi, \lambda) = \prod_{j=1}^{n} f(x_j \mid x_{(j-1)}, s_{(j-1)}; \psi, \lambda) \prod_{j=1}^{n} f(s_j \mid x_{(j)}, s_{(j-1)}; \psi, \lambda); \tag{8}$$

for $j = 1$ the relevant factors are the initial marginal density of x_1 and the conditional density of s_1 given x_1. The partial likelihood of Cox (1975) is the second factor in (8)

$$L_{\rm part}(\psi, \lambda) = \prod_{j=1}^{n} f(s_j \mid x_{(j)}, s_{(j-1)}; \psi, \lambda), \tag{9}$$

and the point of the construction is to find such a sequence so that this part of the full likelihood function depends only on the parameter of interest, with an additional expectation, or hope, that the omitted part does not carry very much information about the parameter of interest. Cox (1975, §2) lists "desiderata" for the construction that is reminiscent of Fisher's (1935) comment:

- : (i) no omitted factor ... should contain important information about the parameters of interest
- : (ii) incidental parameters, and so far as possible nuisance parameters, should not occur in the partial likelihood

The proportional hazards model, as formulated in Cox (1972), has

$$\lambda\{y_i; x_i, \beta, \lambda_0(\cdot)\} = \lambda_0(y_i) \exp(x_i^{\rm T}\beta), \tag{10}$$

where $\lambda(y) = f(y)/\{1-F(y)\}$ is the hazard at time y, i.e. the instantaneous failure rate, given survival up to time y. Given a sample of survival times $y_1 < \cdots < y_n$, along with indicator variables $\delta_1, \ldots, \delta_n$ that record whether or not each y_i was a failure time ($\delta_i = 1$) or a censoring time ($\delta_i = 0$), the joint density under the model (10) is

$$\begin{aligned} L\{\beta, \lambda_0(\cdot)\} &= \prod_{i=1}^{n} \Big(\lambda\{y_i; x_i, \beta, \lambda_0(\cdot)\}[1 - F\{y_i; x_i, \beta, \lambda_0(\cdot)\}] \Big)^{\delta_i} [1 - F\{y_i; x_i, \beta, \lambda_0(\cdot)\}]^{1-\delta_i} \\ &= \prod_{i=1}^{n} \{\lambda_0(y_i) \exp(x_i^T\beta)\}^{\delta_i} \exp\{-\exp(x_i^T\beta)\Lambda_0(y_i)\}. \end{aligned} \tag{11}$$

In (11) $\Lambda_0(y) = \int_0^y \lambda_0(u)du$ is the cumulative baseline hazard function, and x_i is the set of explanatory variables associated with the unit with response time y_i. For simplicity we have assumed the failure times are all distinct, and that the distribution governing the censoring provides no information about either $\lambda_0(\cdot)$ or β.

The partial likelihood for β is much simpler:

$$L_{\rm part}(\beta) = \prod_{failures} \frac{\exp(x_i^T\beta)}{\sum_{k \in R_i} \exp(x_k^T\beta)}, \tag{12}$$

where R_i is the so-called *risk set* at time y_i, i.e. the set of individuals at risk for failure just before the time of the ith failure. The ith term in (12) is the conditional probability that the ith unit fails, given that there was a failure at that time, and given the set of individuals available to fail. The product in (12)

of all the conditional probabilities is not in itself a conditional density, because the conditioning events are overlapping (as they are in (9)); this was the point raised in the discussion by Kalbfleisch and Prentice (1972) and Breslow (1972). The step from (11) to (12) is perhaps not very obvious, and David did allow that it took him some time to arrive at this relatively simple form (Reid, 1994).

However, treating (12) as a regular likelihood function, the analogous approximations of §2 can be described: writing $\ell_{\rm part}(\beta) = \log L_{\rm part}(\beta)$, define the maximum partial likelihood estimate as the solution of $\ell'(\hat\beta_{\rm part}) = 0$, and the partial observed Fisher information $j_{\rm part}(\hat\beta_{\rm part}) = -\ell''_{\rm part}(\hat\beta)$. The usual approximations become

$$\hat\beta_{\rm part} \dot\sim N\{\beta, j^{-1}_{\rm part}(\hat\beta_{\rm part})\}, \quad 2\{\ell_{\rm part}(\hat\beta_{\rm part}) - \ell_{\rm part}(\beta)\} \dot\sim \chi^2_p, \tag{13}$$

the latter under the assumption that $\beta \in \mathbb{R}^p$.

4. Pseudo-likelihood and composite likelihood

A related "construction of convenience" was proposed in Besag (1974) for spatial processes; he called this a pseudo-likelihood function, defined as

$$L_{\rm pseudo}(\theta; y) = \prod_{r=1}^{m} f(y_r \mid y_s; \theta, \text{ site } s \text{ is a neighbour of site } r); \tag{14}$$

reminiscent of an autoregression model in time series. A similar construction of a pseudo-likelihood function was used in Geys et al. (1999), Molenberghs & Verbeke (2006, Ch. 9), and several related papers. For example, Renard et al. (2004) modelled longitudinal binary data with random effects as $\text{pr}(y_{ij} = 1 \mid b_i) = \Phi(x_{ij}^T\beta + z_{ij}^T b_i), j = 1, \ldots, n_i; i = 1, \ldots, m; \quad b_i \sim N(0, \Sigma_b)$, with likelihood function

$$L(\beta, \Sigma_b) = \prod_{i=1}^{m} \int \prod_{j=1}^{n_i} \Phi(x_{ij}^T\beta + z_{ij}^T b_i) db_i, \tag{15}$$

where $\Phi(\cdot)$ is the normal distribution function, and the unobserved random effects, $b_1, \ldots, b_m$, have been integrated out. They proposed a pseudo-likelihood function composed of the joint density of pairs of responses:

$$L_{\rm pseudo}(\beta, \Sigma_b) = \prod_{i=1}^{m} \prod_{r<s} p_{11}^{y_{ir}y_{is}} p_{10}^{y_{ir}(1-y_{is})} p_{01}^{(1-y_{ir})y_{is}} p_{00}^{(1-y_{ir})(1-y_{is})}; \tag{16}$$

each factor representing the probability of one of the four types of pairs $(1,1), (1,0), (0,1)$, and $(0,0)$. The bivariate normal distribution is used to compute p_{00}, p_{01}, etc., and this is computationally more convenient than the calculation of the integrals in (15).

Cox and Reid (2004) considered this pairwise approach as a general construction: given independent response vectors $y_1, \ldots, y_n$, where $y_i \in \mathbb{R}^q$, say, has joint density $f(y_i; \theta)$ they studied inference functions

constructed from marginal and bivariate densities of the components of y_i,

$$\ell_{\text{pair}}(\theta) = \sum_{i=1}^{n} \sum_{s<t} \log\{f_2(y_{is}, y_{it}; \theta)\}, \tag{17}$$

and

$$\ell_{\text{pair}}(\theta) = \sum_{i=1}^{n} \sum_{s<t} \log\{f_2(y_{is}, y_{it}); \theta\} - aq \sum_{i=1}^{n} \log\{f_1(y_{is}; \theta)\}, \tag{18}$$

calling both of these pseudo-likelihoods functions. In later work, (17) has been referred to as a pairwise (marginal) likelihood function. Cox and Reid (2004) considered the formal expansion for the score equation and the second derivative of the log-likelihood function as both q and $n \to \infty$.

Each of partial, pseudo-, and pairwise likelihood functions are examples of a general class of *composite* likelihood functions, so-named, and studied, in Lindsay (1988). Lindsay showed that the estimate obtained by solving the analogue to the score equation, $\ell'_{\text{comp}}(\theta) = 0$, is consistent for θ and asymptotically normally distributed, under some regularity conditions on the model. The asymptotic variance is not given by the negative second derivative of $\ell_{\text{comp}}(\theta)$, but involves the so-called "sandwich formula" for the variance estimate in a misspecified model, derived in a different context in Cox (1961).

Among these various composite likelihood functions, the partial likelihood function (12) is special, in that the asymptotic variance of $\hat{\beta}_{\text{part}}$ **is** indeed estimated by $j^{-1}_{\text{part}}(\hat{\beta}_{\text{part}})$, justifying the use of the first approximation in (13). This was apparently obvious to David, as he writes in Cox (1972, §5): "(16) [j_{part} above] can be used directly for the estimation of variances".

The partial likelihood is also special in a different way: it is connected to semi-parametric likelihoods and empirical likelihoods, suggested by Breslow (1972) for the proportional hazards regression model, and studied more generally in Owen (1988, 2001) and van der Vaart (1998, Ch. 25). Breslow treated the hazard function as constant between observed failure times, the collection of these constants forming the vector of nuisance parameters. Maximizing over these parameters gives a profile likelihood function. In the reply to the discussion, Cox (1972) noted that profile likelihood inference with large numbers of nuisance parameters can be misleading, as mentioned above in §2. In this formulation we have n_0 nuisance parameters, where $n_0 = \Sigma\delta_i$ is the number of distinct failure times. However, in spite of the increasing numbers of nuisance parameters, it is possible to verify that (9) is the profile likelihood function for β, after maximizing over the n_0 constant hazard parameters. This was formally established using a theory of semi-parametric likelihood in Murphy and van der Vaart (2000); the informal argument is sketched in Davison (2003, §10.8). I am grateful to Per Andersen for pointing out that Johansen (1983) set out the argument much earlier, using Aalen's (1978) theory of counting processes; Johansen (1983) noted that the informal argument was developed in an unpublished 1981 thesis by Bay and Mac.

Andersen and Gill (1988) used Aalen's (1978) formulation to prove that the limiting distribution of the estimator $\hat{\beta}_{\text{part}}$ is normal, with asymptotic variance consistently estimated by $j^{-1}_{\text{part}}(\hat{\beta}_{\text{part}})$, and that the limiting distribution of the log-likelihood ratio statistic $2\{\ell_{\text{part}}(\hat{\beta}_{\text{part}}) - \ell_{\text{part}}(\beta)\}$ is χ^2_p, thus justifiying the approximations in (13). In fact $\hat{\beta}_{\text{part}}$ is also asymptotically fully efficient, in the semiparametric model (9).

Other forms of composite likelihood are less special. Because each component of the composite likelihood function is itself a density, $\mathrm{E}_\theta\{\ell'_{\text{comp}}(\theta)\} = 0$, and this is the basis of the argument that the composite maximum likelihood estimator is consistent. However, except in very special cases, the second Bartlett identity fails: $\mathrm{E}_\theta[\{\ell'_{\text{comp}}(\theta)\}^2 + \ell''_{\text{comp}}(\theta)] \neq 0$, so each of these terms must be taken into account in the asymptotic variance of the maximum composite likelihood estimator. For the same reason, the asymptotic distribution of the composite log-likelihood ratio statistic is not χ^2_p, but rather a weighted sum of χ^2_1, with weights related to the two expectations in the equation above. A review of composite likelihood is given in Varin et al. (2011).

The proportional hazards models has been generalized to accommodate data with more complex structure than censored survival data; many generalizations are reviewed in Kalbfleisch and Schnaubel (2023), and partial likelihood has a key role in the analysis of these generalizations. Kalbfleisch and Schnaubel (2023) describe partial likelihood functions for modelling survival data with competing risks and for the analysis of recurrent events; Aalen et al (2008) and Cook and Lawless (2018, 2022) describe partial likelihood functions for event history and other multistate processes. Partial likelihood for changepoint regression models is described in Takeishi (2022).

5. Partial likelihood in practice

Hernandez et al. (2019) present a fairly typical example of the use of Cox's proportional hazards model in practice. Two treatments for septic shock were compared in a multi-center randomized controlled clinical trial, called the ANDROMEDA trial. The response was survival time, censored at 28 days, and the estimate corresponding to the treatment covariate, $\hat{\beta}_1$, say, adjusted for several patient-level covariates, was 0.75, with approximate 95% confidence interval $(0.55, 1.02)$. Because the confidence interval for the hazard ratio includes the value 1, which corresponds to no difference between the treatments, the authors concluded that the new treatment "did not reduce" all-cause mortality, and this led to some online discussion, as the absolute difference in 28-day survival proportions (with no adjustment for covariates, see Table 1) was 43.4% - 34.9% = 8.5%, which in the context of the application is a meaningful clinical difference.

The plot in Figure 1 is in fact the conditional log-likelihood function for the log-odds ratio in Table 1, and it is apparent in the figure that the 95% intervals obtained using either of the approximations in (4) do include the null value 0, but 'just'. Cox (1958a) suggested the use of a confidence distribution function for

	Died	Lived	
New	74	138	212
Old	92	120	212
Total	166	258	424

TABLE 1. Summary data from the ANDROMEDA trial (Hernandez et al., 2019). The full analysis used the proportional hazards model and incorporated patient-level covariates.

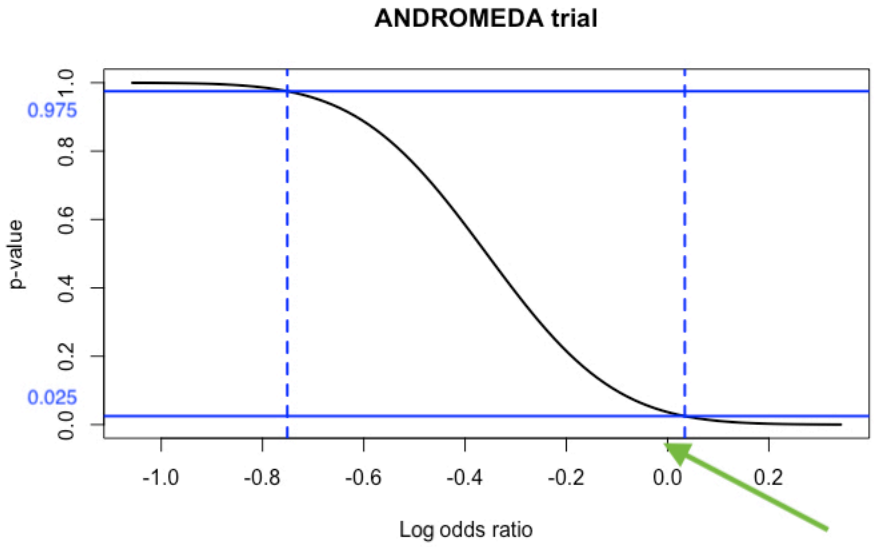

FIGURE 3. Plot of the confidence distribution, or p-value function, for the log-odds ratio, based on the data in Table 1. The 95% confidence interval is indicated by the vertical dashed lines. While the null value 0 is included in this interval (green arrow), the visual impression is that negative values are more compatible with the data.

individual parameters, obtained by varying the confidence probability throughout its range: this function is plotted in Figure 3, with an arrow to the point where the log-odds ratio is 0. Cox (2006, §5.3) wrote "it is tempting to conclude that [the parameter] is more likely to be near the middle of this interval...", while emphasizing that this interpretation of confidence intervals is not correct in the usual frequentist sense.

One way to derive probabilistic statements about the parameter is through a Bayesian approach, and, as part of the discussion of the ANDROMEDA trial, it was proposed that this might be more appropriate. A Bayesian re-analysis of the data was published in Zampieri et al. (2020). In this re-analysis the posterior probability that the odds ratio was less than 1 (favouring the new treatment), was computed for a set of normal priors for the log-odds ratio. The authors reported that this posterior probability ranged from 0.94, for their 'most pessimistic' prior to 0.99, for their 'most optimistic' prior. (This last prior was motivated by the original study design goal of 90% power to detect a reduction from 45% to 30% in 28-day mortality.) It is difficult to compare this directly to the original analysis, as the new analysis is a logistic regression

of the binary outcome "survival past 28 days" (yes/no), and includes a random effect for centre, as well as several fixed effects for the patient-level covariates. One of the priors used for the log odds-ratio was a constant, with the claim that such a flat prior leads to the same inference as a standard non-Bayesian logistic regression of the binary outcome. (This is not strictly correct, as in principle the Bayesian marginal posterior will depend on the choice of priors for the nuisance parameters, but for these data there is little difference.) The conventional logistic analysis leads to a 95% confidence interval that does **not** include the null value, suggesting that if this analysis had been used in the original work, instead of that based on the proportional hazards model, it would have been concluded that the treatment difference was "statistically significant". We might expect that the analysis using the information in the survival times would be more efficient than that using binary responses, so this reversal seems surprising. Steffen Lauritzen in private communication suggested that such a reversal of efficiency could happen if in fact the hazard functions crossed at some time in the first 28 days, so that the proportional hazards assumption was not correct. (A test of the proportional hazards model is provided in the paper, with reported p-value 0.07.).

A Bayesian re-analysis based on an empirical prior derived from published collections of randomized controlled trials, presented in van Zwet et al. (2021), led to an estimated posterior probability of 0.91, somewhat weaker evidence in favour of the new treatment than that based on the priors used in Zampieri et al. (2020).

Bayesian analysis of the proportional hazards model was also used in Naggie et al. (2022), in a platform trial to compare the effect of ivermectin relative to placebo on time to recovery of COVID-19 outpatients with mild or moderate disease. Although the planned analysis was a Bayesian proportional hazards model, the design of the study describes frequentist goals, including the information that the power to detect a hazard ratio of 1.2 was estimated to be 80%. In the statement of findings they write "the posterior probability of improvement in time to recovery in those treated with ivermectin vs placebo had a hazard ratio of 1.07 with a posterior probability of benefit of 0.91. This did not meet the prespecified threshold of posterior probability greater than 0.95. These findings do not support the use of ivermectin in outpatients with mild to moderate COVID-19".

The details of the analysis are somewhat vague, but available software such as SAS Institute Inc. (2023), Wang et al. (2022), or Goodrich et al. (2023), has default settings for the priors for the regression parameters, and for the hazard function, and appears to be based on the full likelihood function (11). In addition to the potential influence of the prior distribution on the marginal posterior, which is particularly risky when there are many nuisance parameters, it is important to emphasize that the partial likelihood, profile likelihood, or any composite likelihood that does not represent the probability of an observable event, cannot be directly combined with a prior distribution to provide a posterior probability; Bayes' rule does not apply. Adjustments

can be constructed to obtain approximately valid inference, as in Pauli et al. (2011), but these adjustments are not straightforward.

A more immediate concern is the public explication of results like the conclusion in Naggie et al. (2022). There has been much ink spilled on the proper interpretation of p-values, and the inadequacy of scientific conclusions that claim an effect is established if "p is less than 0.05", and I agree that this simplistic approach to complex problems is inadequate. At the same time it can hardly be credited that it is an improvement to decide instead to claim that an effect is established if the "posterior probability is greater than 0.95", and it seems to raise a new set of problems in explaining this to a lay public. I am sure David Cox would have concisely and precisely cut through the waffle.

Acknowledgments This paper is based on a talk given at the conference "A Celebration of 50 years of the Cox Model", held at the London School of Hygiene and Tropical Medicine on November 10, 2022. I am grateful to Ruth Keogh and Bianca di Stavola for their efforts in organizing the meeting and this special issue. I thank Heather Battey, Jerry Lawless, and the referee for helpful comments on an earlier version.

Data Availability Statement Data is included in the paper.

Funding Statement The research was partially supported by the Natural Sciences and Engineering Research Council of Canada, Grant Number RGPIN-2020-05897.

References

Aalen, O.O. (1978). Non-parametric inference for a family of counting processes. *Ann. Statist.* **6**, 701–726.

Aalen, O.O., Borgan, Ø. and Gessing, H.K. (2008). *Survival and Event History Analysis: a Process Point of View.* Springer, New York.

Andersen, P.K. and Gill, R.D. (1988). Cox's regression model for counting processes: a large-sample study. *Ann. Statist.* **10**, 1100–1120.

Battey, H.S. (2024). D.R. Cox: Aspects of scientific inference. *J. R. Statist. Soc* A, to appear.

Battey, H.S., Cox, D.R. and Lee, S. (2023). On partial likelihood and the construction of factorisable transformations. *Information Geometry* **7**, 9–28.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. R. Statist. Soc.* B **36**, 192–236.

Breslow, N. (1972). Discussion of Regression models and life-tables by D.R. Cox. *J. R. Statist. Soc.* B **34**, 216–217.

Cook, R.J. and Lawless, J.F. (2018). *Multistate Models for the Analysis of Life History Data.* CRC/Chapman and Hall, Boca Raton.

Cook, R.J. and Lawless, J.F. (2022). Life history analysis with multistate models: a review and some current issues. *Canad. J. Statist.* **50**, 1270–1298.

Cox, D.R. (1958a). Some problems connected with statistical inference. *Ann. Math. Statist.* **29**, 357–372.

Cox, D.R. (1958b). The regression analysis of binary sequences (with discussion). *J. R. Statist. Soc.* B **20**, 215–242.

Cox, D.R. (1961). Tests of separate families of hypotheses. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, **1**, edited by L.M. LeCam, J. Neyman and E.L. Scott. University of California Press, Berkeley, 105–123.

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

Cox, D.R. (1975). Partial likelihood. *Biometrika* **62**, 269–276.

Cox, D.R. and Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika* **91**, 729–737.

Cox, D.R. (2006). *Principles of Statistical Inference*. Cambridge University Press, Cambridge.

Davison, A.C. (2003). *Statistical Models*. Cambridge University Press, Cambridge.

Fisher, R.A. (1935). The logic of inductive inference. *J. R. Statist. Soc.* **98**, 39–82.

Geys, H., Molenberghs, G. and Ryan, L.M. (1999) Pseudolikelihood modelling of multivariate outcomes in developmental toxicology. *J. Am. Statist. Assoc.* **94**, 734–745.

Goodrich B., Gabry J., Ali I., and Brilleman S. (2023). `rstanarm`: Bayesian applied regression modeling via Stan. R package version 2.21.4. `https://mc-stan.org/rstanarm`.

Hernández, G., Ospina-Tascón, G.A. et al. (2019). Effect of a resuscitation strategy targeting peripheral perfusion status versus serum lactate levels on 28-day mortality among patients with septic shock: the ANDROMEDA-SHOCK randomized clinical trial. *J. Am. Med. Assoc.* **321**, 654–664.

Johansen, S. (1983). An extension of Cox's regression model. *Inter. Statist. Rev.* **51**, 165–174.

Kalbfleisch, J.D. and Prentice, R.L. (1972). Discussion of Regression models and life-tables by D.R. Cox. *J. R. Statist. Soc.* B **34**, 215–216.

Kalbfleisch, J.D. and Schnaubel, D.E. (2023). Fifty years of the Cox model. *Ann. Rev. Statist. Applic.* **10**, 1–23.

Lindsay, B.G. (1988). Composite likelihood methods. *Contemp. Math.* **80**, 220–239.

Molenberghs, G. and Verbeke, G. (2006). *Models for Discrete Longitudinal Data*. Springer, New York.

Murphy, S.A. and van der Vaart, A. (2000). On profile likelihood. *J. Am. Statist. Assoc.* **95**, 449–465.

Naggie, S., Boulware, D.R., et al. (2022). Effect of ivermectin vs placebo on time to sustained recovery in outpatients with mild to moderate COVID-19. *J. Am. Med. Assoc.* **328**, 1595–1603.

Neyman, J. and Scott, E.L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16**, 1–32.

Owen, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237–249.

Owen, A.B. (2001). *Empirical Likelihood*. CRC/Chapman and Hall, London.

Pauli, F., Racugno, W. and Ventura, L. (2011). Bayesian composite marginal likelihoods. *Statist. Sinica* **21**, 149–164.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Reid, N. (1994). A conversation with Sir David Cox. *Statist. Sci.* **9**, 439–455.

Renard, D., Molenberghs, G. and Geys, H. (2004). A pairwise likelihood approach to estimation in multilevel probit models. *Comput. Statist. Data Anal.* **44**, 649–667.

SAS Institute Inc. (2023). "The PHREG Procedure" in SAS/STAT 15.3 User's Guide. SAS Institute Inc., Cary NC.

Takeishi, S. (2022). Asymptotic properties of the smoothed partial likelihood estimator in the change-plane Cox model. *Scand. J. Statist.* **50**, 1503–1531.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

van Zwet, E., Schwab, S. and Senn, S. (2021). The statistical properties of RCT's and a proposal for shrinkage. *Statist. Med.* **40**, 6107–6117.

Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods. *Statist. Sinica* **21**, 5–42.

Wang W., Chen M., Wang X., Yan J. (2022). `dynsurv`: Dynamic models for survival data. R package version 0.4-3. `https://CRAN.R-project.org/package=dynsurv`

Zampieri, F.G., Damiani, L.P., et al. (2020). Effects of a resuscitation strategy targeting peripheral perfusion status versus serum lactate levels among patients with septic shock: A Bayesian reanalyis of the ANDROMEDA-SHOCK trial. *Am. J. Respir. Crit. Care Med.* **201**, 423–429.