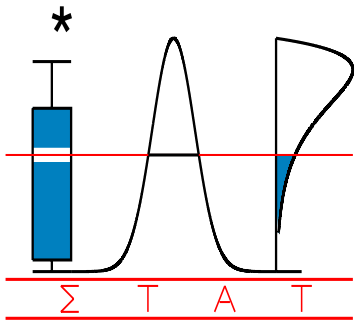
T E C H N I C A L

R E P O R T

0344

PENALIZED PARTIAL LIKELIHOOD FOR FRAILTIES AND SMOOTHING SPLINES IN TIME OF FIRST ISEMINATION MODELS FOR DAIRY COWS

Luc DUCHATEAU and Paul JANSSEN



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Penalized partial likelihood for frailties and smoothing splines in time to first insemination models for dairy cows

Luc Duchateau

Department of Physiology, Biochemistry and Biometrics, Faculty of Veterinary Medicine, Ghent University, Salisburylaan 133, 9820 Merelbeke, Belgium.

email: Luc.Duchateau@Ugent.be

and

Paul Janssen

Limburgs Universitair Centrum, Center for Statistics, Universitaire Campus, 3590 Diepenbeek, Belgium.

SUMMARY. In many epidemiological studies time to event data are clustered and the physiological relationship between (time dependent) covariates and the log hazard is often not linear as assumed in the Cox model. Introducing frailties in the Cox model can account for the clustering of the data and smoothing splines can be used to describe nonlinear relations. These two extensions of the Cox model are introduced jointly and it is shown how penalized partial likelihood techniques can be used to fit the extended model. We demonstrate the need for such a model to study the relation between the physiological covariates milk ureum and protein

concentration and the log hazard of first insemination in dairy cows, with the farms as clusters.

KEY WORDS: Penalized partial likelihood; frailty model; smoothing splines; time to first insemination; dairy cows.

1 Introduction

One of the essential variables in maximizing the milk production in a dairy farm is the intercalving time (Dijkhuizen et al., 1997). A long intercalving time corresponds to low milk production at the end of the lactation. Since the widespread use of artificial insemination, the time to first insemination has been recognized as the most important variable influencing the intercalving time of a cow (Ferguson and Galligan, 1999). A cow is inseminated when it shows signs of being receptive (often based on heat detection). It is thus important to study whether particular milk parameters that are followed up on a regular basis can be used to predict time to first insemination in order to track down fertility problems so that remedial measures can be taken. Two such covariates, the milk ureum and protein concentration, are investigated here.

A standard way to study the relation between the time to first insemination (possibly censored) and milk ureum or protein concentration (time dependent covariates) is to model the log hazard as the sum of a baseline log hazard and a linear term in the covariate; this is the Cox model (Cox, 1972). The assumption that the log

hazard depends on the covariate in a linear way is often not true when investigating physiological phenomena. It is therefore important to test the linearity assumption and, if instrumental, to allow for a more complex dependence of the log hazard on the covariate through a function $g(.)$. One such approach is to define $g(.)$ as a smoothing spline. A further extension of the Cox model is obtained by taking into account the hierarchical (clustered) structure of the data, e.g., in our study dairy cows are nested within farms. This clustering can be taken into account by adding a random effect as extra term. Note that for $g(x) = \beta_0 x$, the latter model is the shared frailty model (Klein, 1992). The farm is taken to be a random effect rather than a fixed effect because the individual farm is not of interest by itself; interest is rather in the heterogeneity between farms. Furthermore, introducing many fixed effects in a model might lead to convergence problems, especially if there is little variation in the covariates between farms (McGilchrist and Aisbett, 1991).

An estimate of the risk parameter (the regression coefficient β_0) in the Cox model is the maximizer of the partial likelihood. To estimate the parameters in the Cox model extended with the smoothing spline, penalized partial likelihood methodology is used as estimation tool (Gray, 1992). For the shared frailty model with a gamma distributed frailty, estimates of the parameters can be obtained by the use of the EM algorithm. An alternative and elegant estimation tool is penalized partial likelihood maximization. See Therneau et al. (2003) for a review and Therneau and Grambsch (2000) for a detailed discussion.

In this paper we consider the extension of the Cox model where the dependence

of the log hazard on the covariate is modeled as a spline function and where we allow for frailties. It will be discussed how the parameters in this model can be estimated using penalized partial likelihood methods. Details on specific penalized partial likelihood features for frailties and smoothing splines will be given.

For the study on insemination in dairy cows it is shown that it is essential to include frailties and smoothing splines in one single model to arrive at a valid statistical analysis of the data.

Finally, note that penalized partial likelihood methods have a long history in models with normally distributed error terms: see Henderson (1953) for hierarchical models and Wegman and Wright (1983) for smoothing splines. Penalized partial likelihood methods for (censored) survival data are more recent (Gray, 1992).

2 The ordinary Cox model and its two extensions

We first introduce the ordinary Cox model with the corresponding partial likelihood for our data example. The data consist of cows clustered in different farms. A cow is followed-up for several time varying milk covariates and the objective is to model the time to first insemination as a function of these time varying covariates.

The most appropriate model to describe time to event data with time varying covariates is the Cox model

$$h_{ij}(t) = h_0(t) \exp\{\beta_0 x_{ij}(t)\} \tag{1}$$

where $h_{ij}(t)$ is the hazard rate at time t for cow j from farm i, $h_0(t)$ is the unspecified baseline hazard at time t, $x_{ij}(t)$ is the value for the covariate at time t for cow j from farm i and β_0 is the linear change of the log hazard rate with one unit change in the covariate.

The relevant information for an individual cow j ($j = 1, \ldots, n_i$) from farm i ($i = 1, \ldots, n$) is contained in the vector

$$\{t_{ij}, \delta_{ij}, x_{ij}(t_{ij1}), \ldots, x_{ij}(t_{ijl_{ij}})\}$$

with t_{ij} the time to first insemination or censoring, δ_{ij} the censoring indicator and $x_{ij}(t_{ij1}), \ldots, x_{ij}(t_{ijl_{ij}})$ the values for the covariates recorded at times $t_{ij1}, \ldots, t_{ijl_{ij}}$. As the covariate is only determined once a month, its value at a particular time-point t, $x_{ij}(t)$, is determined by linear interpolation based on the measurements immediately before and after time t.

An estimate for the parameter of interest, β_0, can be found by maximizing the partial likelihood

$$\prod_{i=1}^{n} \prod_{j=1}^{n_i} \left[\frac{\exp\{\beta_0 \; x_{ij}(t_{ij})\}}{\sum_{R(t_{ij})} \exp\{\beta_0 x_{qs}(t_{ij})\}} \right]^{\delta_{ij}}$$

where $R(t_{ij})$ denotes summation over all (q, s) indices for which $t_{qs} \geq t_{ij}$, i.e. the sum over all animals in the risk set at time t_{ij}.

Up to now, it has been assumed that the effect of the time varying covariates on the hazard of first insemination is linear on the log scale. In practice, however,

such relationships are often not linear and therefore it is worthwhile to test for nonlinearity and, in the case of nonlinearity, to be able to depict the evolution of the hazard of first insemination as a function of the milk ureum or protein concentration. Thus the linear relationship of the time varying covariate, described by $\beta_0 x_{ij}(t)$ is replaced by a flexible function $g\{x_{ij}(t)\}$ resulting in

$$h_{ij}(t) = h_0(t) \exp\left[g\{x_{ij}(t)\}\right]. \tag{2}$$

In section 3, one particular model with such a flexible function approach that can be fitted by penalized partial likelihood will be given.

A further extension of the Cox model is needed as cows are clustered within farms. The clustering can be accounted for by adding the farm as a frailty factor leading to

$$h_{ij}(t) = h_0(t) u_i \exp\{\beta_0 x_{ij}(t)\} \tag{3}$$

where u_i is the frailty for farm i assumed to be a realization of a frailty density. Here the one parameter gamma density

$$f_U(u) = \frac{u^{\frac{1}{\theta}-1} \exp(\frac{-u}{\theta})}{\theta^{\frac{1}{\theta}} \Gamma(\frac{1}{\theta})}$$

is used for its mathematical convenience.

This extended model thus contains one more parameter, θ, the variance of the frailties, describing the heterogeneity between farms. In section 4, the penalized partial likelihood approach to fit this model will be described in detail.

The final model combines these two extensions in one model

$$h_{ij}(t) = h_0(t)u_i \exp\left[g\{x_{ij}(t)\}\right]. \tag{4}$$

It will be shown in section 5 how penalized partial likelihood can be used to fit this model; and in section 6 the relevance of this model will be demonstrated for the milk production case study.

3 Penalized partial likelihood for smoothing splines

As discussed in Sleeper and Harrington (1990) and Gray (1992) a spline function is a natural choice for approximating the covariate transformation $g\{x_{ij}(t)\}$. More precisely, with $B_1(x), \ldots, B_{K+4}(x)$ the cubic B-spline basis for the space of the cubic splines with K prespecified knots, we take

$$g(x) = \beta_0 x + \sum_{k=1}^{K+2} \beta_k B_k(x).$$

We only include $K+2$ (of the $K+4$) basis splines because the constant term can be absorbed in the baseline hazard and because the linear term is specified separately. Any of the two B-spline functions can be dropped, provided the resulting parameterization is of full rank.

Sleeper and Harrington (1990) use the partial likelihood to estimate the parameters. Since often interest is in alternatives that deviate from the linear term in a smooth way, Gray (1992) subtracts the penalty term $\lambda \int \{g''(z)\}^2 dz$ (λ times the integrated curvature of g), i.e., he considers the penalized partial likelihood

$$\ell^{\mathrm{S}}_{\mathrm{ppl}}(\boldsymbol{\beta}) = \log \left[\prod_{i=1}^{n} \prod_{j=1}^{n_i} \left[\frac{\exp\left[g\{x_{ij}(t_{ij})\}\right]}{\sum_{R(t_{ij})} \exp\left[g\{x_{qs}(t_{ij})\}\right]} \right]^{\delta_{ij}} \right] - \lambda \int \{g''(z)\}^2 dz \qquad (5)$$

with $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{K+2})$ the parameters defining the function g.

The idea is that we decrease the likelihood by subtracting the term that accounts for the roughness of the function g. The extra factor λ is a tuning parameter for the penalty we impose; it governs the trade-off between the likelihood term and the penalty term. For a fixed value of the smoothing parameter λ the penalized partial likelihood can be maximized to obtain parameter estimates.

The smoothing parameter λ can be selected by the user, but it is more appropriate to rely on methods that automatically select the value of the smoothing parameter, such as cross-validation (Verweij and van Houwelingen, 1993) or the minimisation of Akaike's Information Criterion (AIC) (Akaike, 1973). A corrected AIC criterion (Hurvich et al., 1998) is used in the example to determine the smoothing parameter.

4 Penalized partial likelihood for frailty models

The addition of frailties to the Cox model leads to unobserved entities in the model which also prevail in the partial likelihood. It is however assumed that these frailties

come from a gamma density with mean equal to 1 and unknown heterogeneity parameter θ. Therefore, a penalty is added to the partial likelihood that decreases with the distance of the frailty from one, the mean of the frailty density.

The penalty term on the log scale in the case of the gamma density is given by

$$-\sum_{i=1}^{n} \log\{f_U(u_i)\}.$$

The penalized partial likelihood for the frailty model is then given (McGilchrist, 1993) by

$$\ell_{\text{ppl}}^{\text{F}}(\beta_0, \boldsymbol{u}, \theta) = \log\left[\prod_{i=1}^{n}\prod_{j=1}^{n_i}\left[\frac{u_i \exp\{\beta_0\ x_{ij}(t_{ij})\}}{\sum_{R(t_{ij})} u_q \exp\{\beta_0 x_{qs}(t_{ij})\}}\right]^{\delta_{ij}}\right] + \sum_{i=1}^{n} \log\{f_U(u_i)\}. \tag{6}$$

For fixed values of the heterogeneity parameter θ, maximization of the penalized partial likelihood criterion leads to the same parameter estimates for the fixed effects β and the frailties u_i as the EM-algorithm (Therneau et al., 2003). For a particular value of θ, estimates for the fixed effects, frailties and baseline hazards can thus be obtained by maximizing the penalized partial likelihood.

To make clear that we keep θ fixed in $\ell_{\text{ppl}}^{\text{F}}(\beta_0, \boldsymbol{u}, \theta)$, we write $\ell_{\text{ppl}}^{\text{F}}(\beta_0, \boldsymbol{u} \mid \theta)$; we further use $\hat{\beta}_0^{\theta}$ and $\hat{\boldsymbol{u}}^{\theta}$ to denote the values of β_0 and $\boldsymbol{u}$ that maximize, for the given value of θ, $\ell_{\text{ppl}}^{\text{F}}(\beta_0, \boldsymbol{u} \mid \theta)$. We now consider the profile penalized partial likelihood $\ell_{\text{ppl}}^{\text{F}}(\hat{\beta}_0^{\theta}, \hat{\boldsymbol{u}}^{\theta} \mid \theta)$ as a function of θ. From the discussion in section 6 it will be seen that the estimate of θ obtained from the EM-algorithm can not be obtained by

maximizing the profile penalized partial likelihood. The profile penalized partial likelihood is indeed increasing with increasing values of θ.

A way to obtain an estimate for θ that corresponds with the EM-estimate is to replace the profile penalized partial likelihood by the profile marginal likelihood of θ and to estimate θ as the argument that maximizes this profile marginal likelihood. The marginal likelihood is obtained by integrating out the frailties from the joint density of the observed event/censoring time and the frailties (Klein, 1992) and is of form

$$\ell^{\mathrm{F}}_{\mathrm{marg}}\left(\theta, h_0(.), \beta_0\right) = \sum_{i=1}^{n} \left[D_i \log\theta - \log\Gamma\left(1/\theta\right) + \log\Gamma\left(1/\theta + D_i\right) \right.$$
$$\left. - \left(1/\theta + D_i\right) \log\left[1 + \theta \sum_{j=1}^{n_i} H_{ij}(t_{ij})\right] + \sum_{j=1}^{n_i} \delta_{ij}\{\beta_0 x_{ij}(t_{ij}) + \log h_0(t_{ij})\} \right]$$

with D_i the number of events at farm i and $H_{ij}(.)$ the cumulative hazard for animal j in farm i.

To arrive at the profile marginal likelihood we replace β_0, $h_0(.)$ and $H_{ij}(.)$ in this general expression for the marginal likelihood by their estimates $\hat{\beta}_0^\theta$, $\hat{h}_0^\theta(.)$ and $\hat{H}_{ij}^\theta(.)$. In terms of the estimates $\hat{\beta}_0^\theta$ and $\hat{\boldsymbol{u}}^\theta$ we can give explicit expressions for the estimated baseline hazard and cumulative baseline hazard. With e the total number of ordered distinct event times $t_{(1)} < \ldots < t_{(e)}$ and with $d_{(k)}$ the number of events at time $t_{(k)}, k = 1, \ldots, e$, define (as in Duchateau et al. (2002))

$$\hat{h}_0^\theta(t_{(k)}) = \frac{d_{(k)}}{\sum\limits_{R(t_{(k)})} \hat{u}_q^\theta \exp\{\hat{\beta}_0^\theta x_{qs}\left(t_{(k)}\right)\}}. \tag{7}$$

An estimate for the cumulative hazard $H_{ij}(t_{ij})$ is

$$\hat{H}_{ij}^\theta(t_{ij}) = \sum_{t_{(l)} \leq t_{ij}} \hat{h}_0^\theta(t_{(l)}) \exp\{\beta_0 x_{ij}(t_{(l)})\}.$$

The approach described above has been implemented in the Splus function 'coxph' (Therneau and Grambsch, 2000).

Remark. The penalty term for the frailty model and the smoothing splines resemble each other. The penalty term for the smoothing splines consists of a tuning parameter and the parameters that define the function $g(x)$; that vector $\boldsymbol{\beta}$ also occurs in the partial likelihood part. Similarly, the penalty term for the frailty model contains the parameter θ and the frailties which also occur in the partial likelihood part. Neither the tuning parameter nor θ occur in the partial likelihood term. As compared however with the penalized partial likelihood for smoothing splines, where λ can be chosen by the user, this is not the case for the frailty model as θ itself is also a parameter that needs to be estimated from the data.

5 Smoothing splines and frailty models combined

In the two previous sections, smoothing splines and frailties in the context of the Cox model are dealt with separately. It is necessary when studying nonlinear relationships in clustered survival data to combine smoothing splines and frailties as

presented in model (4). For such models parameter estimates can be obtained by an iterative procedure. The following conditional likelihoods are needed in the iterative procedure.

The penalized partial likelihood for smoothing splines as in (5), but with the frailties $\boldsymbol{u}$ as fixed offset term

$$\ell_{\text{ppl}}^{\text{S}}(\boldsymbol{\beta} \mid \boldsymbol{u}) = \log\left[\prod_{i=1}^{n}\prod_{j=1}^{n_i}\left[\frac{u_i \exp\left[g\{x_{ij}(t_{ij})\}\right]}{\sum\limits_{R(t_{ij})} u_q \exp\left[g\{x_{qs}(t_{ij})\}\right]}\right]^{\delta_{ij}}\right] - \lambda\int\{g''(z)\}^2 dz \tag{8}$$

and the penalized partial likelihood for a particular value of θ as in (6), but with $g(x)$ as fixed offset term

$$\ell_{\text{ppl}}^{\text{F}}(\boldsymbol{u} \mid \theta, \boldsymbol{\beta}) = \log\left[\prod_{i=1}^{n}\prod_{j=1}^{n_i}\left[\frac{u_i \exp\left[g\{x_{ij}(t_{ij})\}\right]}{\sum\limits_{R(t_{ij})} u_q \exp\left[g\{x_{qs}(t_{ij})\}\right]}\right]^{\delta_{ij}}\right] + \sum_{i=1}^{n}\log\{f_U(u_i)\}. \tag{9}$$

Finally, we will use the marginal likelihood with $g(x)$ and $h_0(.)$ as fixed offset term

$$\begin{aligned}\ell_{\text{marg}}^{\text{F}}(\theta, h_0(.), \boldsymbol{\beta}) = \sum_{i=1}^{n}\Bigg[&D_i \log\theta - \log\Gamma\left(1/\theta\right) + \log\Gamma\left(1/\theta + D_i\right) \\ &- \left(1/\theta + D_i\right)\log\left[1 + \theta\sum_{j=1}^{n_i} H_{ij}(t_{ij})\right] + \sum_{j=1}^{n_i}\delta_{ij}\left[g\{x_{ij}(t_{ij})\} + \log h_0(t_{ij})\right]\Bigg].\end{aligned} \tag{10}$$

The iterative procedure based on these conditional likelihoods then goes as follows:

1. Initialize $\boldsymbol{u}^1 = (u_1^1, \ldots, u_n^1)$ with $u_i^1 = 1, i = 1, \ldots, n$ and set $r = 1$ (r counts the iteration steps).

2. Maximize the penalized partial likelihood (8) inserting $\boldsymbol{u}^r$ as fixed offset term and obtain $\boldsymbol{\beta}^r$ for the smoothing splines, i.e., $\boldsymbol{\beta}^r = \arg\max \ell_{\text{ppl}}^{\text{S}}(\boldsymbol{\beta} \mid \boldsymbol{u}^r)$. Insert these values in $g(x)$ to obtain the estimate $g^r(x)$ at the r^{th} iteration step.

3. Now use $g^r(x)$ as fixed offset term to find estimates for θ and $\boldsymbol{u}$. This step itself consists of an outer and an inner loop.

   Outer loop Take a grid G of θ values.

   Inner loop

   - Take $g^r(x)$ (or equivalently $\boldsymbol{\beta}^r$) and fix a θ from G.
   - Obtain $\boldsymbol{u}^{\theta,r} = \arg\max \ell_{\text{ppl}}^{\text{F}}(\boldsymbol{u} \mid \theta, \boldsymbol{\beta}^r)$.
   - Use $\boldsymbol{u}^{\theta,r}$ and $\boldsymbol{\beta}^r$ to obtain $h_0^{\theta,r}(.)$ from (7) now using $g^r(x)$ rather than $\beta_0 x$.
   - Calculate $\ell_{\text{marg}}^{\text{F}}\left(\theta, h_0^{\theta,r}(.), \boldsymbol{\beta}^r\right)$.

   Let $\theta^r = \arg\max_{\theta \in G} \ell_{\text{marg}}^{\text{F}}\left(\theta, h_0^{\theta,r}(.), \boldsymbol{\beta}^r\right)$ and let $\boldsymbol{u}^{r+1} = \arg\max \ell_{\text{ppl}}^{\text{F}}(\boldsymbol{u} \mid \theta^r, \boldsymbol{\beta}^r)$.

4. Check whether the algorithm has converged. If not, increase the iteration step with 1, $r = r + 1$ and go back to step 2.

Note. In practice, the grid search is replaced by a golden section search.

6 Example

6.1 Data

The data used for this study are extracted from the database of the regional Dairy Herd Improvement Association which includes the official recording system and artificial insemination (AI) service programme. In total, 7973 cows from 181 dairy farms are considered. The median number of cows per farm equals 41 cows. Days to first insemination from parturition for those cows that were inseminated were recorded. No first insemination date was recorded for 1287 cows which were considered to be censored at the last follow-up day. Furthermore, milk protein and ureum concentrations were determined monthly. We investigate the effect of these two covariates separately because the objective is to assess whether any of these two covariates can be used to predict a delay in time to first insemination and thus to detect fertility problems. The extension of the methods described above to the case of more covariates is straightforward.

6.2 Why frailties: the ureum concentration

A frailty model with smoothing splines for the time varying ureum concentration is fitted to the time to first insemination data. The test for nonlinearity iss not significant (p=0.43), so that only the linear effect of the ureum concentration on the log hazard of first insemination is further considered.

The estimated hazard ratio within one farm is equal to 0.949 (95% CI: [0.915;0.985])

with lower ureum concentration having significantly higher hazard of first insemination (p=0.0057). The heterogeneity parameter θ is equal to 0.334. Fitting a basic Cox regression model without frailties for the farm effect leads to a hazard ratio equal to 0.973 (95% CI: [0.943;1.005]) with no significant effect of ureum concentration on the hazard of first insemination (p=0.094).

In order to understand the reason for the different results of the two models, the log hazard ratio, β_i, is estimated separately for each farm and depicted as a function of the log of the predicted farm frailty, $\log(\hat{u}_i)$ (Figure 1a). There are several farms for which the frailty is substantially smaller than the mean of the frailty density function. The log hazard ratio of 6 of these farms is exactly equal to 0 as no inseminations at all were taking place in these farms. Therefore the log hazard ratio is estimated to be 0 (since all $\delta_{ij} = 0$) as obviously no relationship can be found in the case no events take place.

For each of the farms, the mean ureum concentration is calculated as

$$\bar{x}_i = \frac{\sum_{j=1}^{n_i} \sum_{k=1}^{l_{ij}} x_{ij}(t_{ijk})}{\sum_{j=1}^{n_i} l_{ij}}, \quad i = 1, \ldots, n$$

and compared to the overall mean ureum concentration

$$\bar{x} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n_i} \sum_{k=1}^{l_{ij}} x_{ij}(t_{ijk})}{\sum_{i=1}^{n} \sum_{j=1}^{n_i} l_{ij}}.$$

The 6 farms without any inseminations have either average values for the mean ureum concentration, or are below the average overall ureum concentration (Figure 1b). Both the simple Cox model and the frailty model predict that cows with low

ureum concentration have higher hazard of first insemination but the farms with no inseminations at all and low to average mean ureum concentration contradict this global relationship. The frailty model corrects for this by assigning a large negative value for the log frailty and thus all cows of these farms have low hazard. This obviously does not happen in the simple Cox model without frailties, so that these cows will contradict the overall relationship between ureum concentration and hazard of first insemination. Because of this, the overall hazard ratio in the basic Cox model is closer to 1 and no longer significant.

We now study the behavior of the two terms of the penalized partial likelihood separately. As discussed in section 4, the profile penalized partial likelihood for θ increases with increasing values of θ (Figure 2). Obviously, the partial likelihood part alone increases and the penalty term increases with increasing values of θ. For small values of θ, the penalty term becomes negative leading to a larger value for the penalized partial likelihood than for the partial likelihood term alone. This is due to the fact that all frailty terms are close to the mean of a density function with small variance, and therefore most of the contributions from the density function are larger than 1 and the logarithm thus positive.

Secondly, we further study the behavior of the profile marginal likelihood for θ, for which the maximum leads to the estimate of θ. Statistical inference thus needs to be based on the marginal likelihood. Andersen et al. (1997) determined the variance of θ using the Hessian of the marginal likelihood including as parameters the functions $\hat{h}_0(.)$ as defined in (7). For large datasets as in our case this procedure is

computationally intensive as the number of parameters corresponds to the number of different event times plus the number of fixed effects plus 1 for θ. Furthermore, the variance is in many cases not a useful parameter to derive confidence intervals as especially in the neighborhood of 0 the marginal likelihood can be rather skewed. Therefore, we believe it is a better procedure to use the profile marginal likelihood to determine confidence intervals. For instance, when approximating the profile marginal likelihood by χ^2_1, we need to take those two values of θ for which the marginal profile likelihood lies 1.92 units below the maximum profile likelihood value for the 95% confidence interval (Morgan, 1992). In our data example, this corresponds to the interval [0.253;0.410] (see Figure 3).

6.3 Why smoothing splines: the protein concentration

The frailty model with time varying milk protein concentration as a linear fixed effect leads to an estimated hazard ratio equal to 1.51 (95% CI: [1.39;1.65]), which is significantly different from 1 ($p<0.0001$) and θ is estimated to be 0.319. The non-linear terms, however, are also significantly different from 0 at the log hazard scale ($p<0.0001$) and can thus not be eliminated from the model. The cubic spline relationship between the milk protein concentration and the log hazard function is shown in Figure 4 for three different values of λ ($\lambda = 0.2, 0.4, 0.725$) with $\lambda = 0.725$ the smoothing parameter value that maximizes the AIC. The log hazard of a particular protein concentration x is expressed relative to the mean protein concentration $\bar{x}$ over all protein concentration measurements of the different cows in the different

farms

$$\log\{\frac{h(x)}{h(\bar{x})}\} = \log\{h(x)\} - \log\{h(\bar{x})\}$$

Whatever the value of λ, a linear increase is apparent in the range of the lower milk protein concentrations below 3.2 %. For higher milk protein concentrations, however, the linear relationship no longer holds. There seems to be an optimal milk protein concentration value around 3.6%. For lower and higher protein concentrations, the hazard decreases. When fitting only a linear relationship, a large hazard ratio significantly different from 1 is found. This is due to the fact that most milk protein concentrations measured are on the low side and in the range where the linear relationship holds. As can be read from the left panel of Figure 4, 90% of the observed protein concentrations were below 3.6%.

7 Conclusions

When investigating physiological relationships in clustered time to event data, it is important to model the clustering and simultaneously allow for flexible, non-linear effects of the covariates on the log hazard. This can be done in the frailty model framework with smoothing splines. The model fitting is partially based on the maximization of the penalized partial likelihood. However, penalized partial likelihood in the context of the frailty model can not be used to find estimates of θ, the variance of the frailties. The penalized partial likelihood is rather a technique to find estimates for the fixed effects and frailties given a particular value of θ. Instead,

estimation of θ is based on the profile marginal likelihood. Furthermore, profiling the marginal likelihood for θ is also an easy and adequate technique to derive the 95% confidence interval for θ.

Acknowledgements

The authors acknowledge Geert Opsomer for providing the insemination data and the referees for useful comments and remarks. The research of Paul Janssen has been supported by the Ministry of the Flemish Community (Project BIL00/28, International Scientific and Technological Cooperation) and the Interuniversity Attraction Poles research network P5/24 of the Belgian State (Federal Office for Scientific, Technical and Cultural Affairs).

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Inference Theory*, B. N. Petrov and Csáki F. (eds), 267-281. Budapest: Akadémiai Kiadó.

Andersen, P. K. , Klein, J. P. , Knudsen, K. M. and Tabanera y Palacios, R. (1997). Estimation of variance in Cox's regression model with shared gamma frailties. *Biometrics* **53**, 1475-1484.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187-220.

Dijkhuizen, A. A., Huirne, R. B. M., Jalvingh, A. W. and Stelwagen, J. (1997). Economic impact of common health and fertility problems. In *Animal health economics, principles and applications*, A.A. Dijkhuizen and R. S. Moore (eds), 41-58. University of Sydney.

Duchateau, L., Janssen, P., Lindsey, P., Legrand, C., Nguti, R., Sylvester, R., 2002. The shared frailty model and the power for heterogeneity tests in multicenter trials. *Computational Statistics and Data Analysis* **40**, 603-620.

Ferguson, J. D. and Galligan, D. T. (1999). Veterinary reproductive programs. In *Proceedings of the 32nd Annual Conference of the American Association of Bovine Practitioners*, 131-137. Nashville, Tenessee.

Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association* **87**, 942-951.

Henderson C. R. (1953). Estimation of variance and covariance components. *Biometrics* **9**, 226-252.

Hurvich, C. M. , Simonoff, J. S. and Tsai, C. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society, Series B* **60**, 271-293.

Klein, J. P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics* **48**, 795-806.

McGilchrist, C. A. and Aisbett, C. W. (1991). Regression with frailty in survival analysis. *Biometrics* **47**, 461-466.

McGilchrist, C. A. (1993). REML estimation for survival models with frailty. *Biometrics* **49**, 221-225.

Morgan, B. J. T. (1992). *Analysis of quantal response data*, 63-67. London: Chapman and Hall.

Sleeper, L. A. and Harrington, D. P. (1990). Regression splines in the Cox model with application to covariate effects in liver disease. *Journal of the American Statistical Association* **85**, 941-949.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling survival data. Extending the Cox model.* Springer Verlag, New York.

Therneau, T. M. , Grambsch, P. M. and Pankratz, V. S. (2003). Penalized survival models and frailty. *Journal of Computational and Graphical Statistics* **12**, 156-175.

Verweij, P. J. M. and van Houwelingen, H. C. (1993). Cross-validation in survival analysis. *Statistics in Medicine* **12**, 2305-2314.

Wegman, E. J. and Wright, I. W. (1983). Splines in statistics. *Journal of the American Statistical Association* **78**, 351-365.

Figure 1: The relationship between the log of the predicted farm frailty and the within farm log hazard ratio for the ureum concentration (a) and the mean farm ureum concentration (b). The dashed line corresponds to the overall mean ureum concentration. Triangles represent farms with log hazard ratio equal to 0.

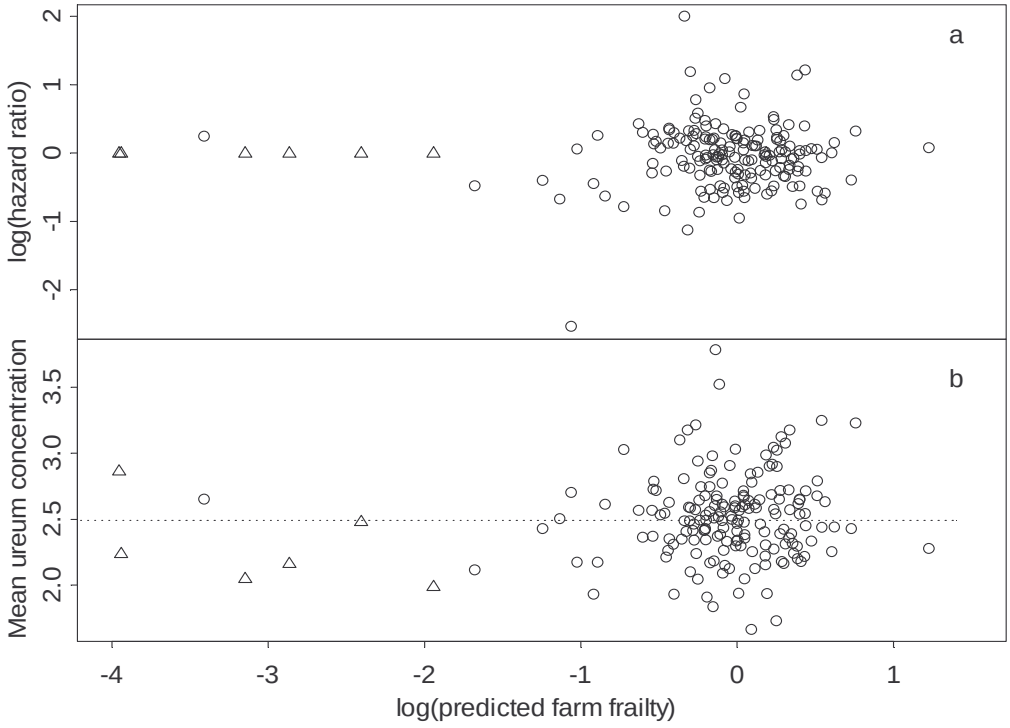


Figure 2: Profile penalized partial likelihood for θ, with the solid line the penalized partial likelihood, the dashed line the partial likelihood part alone in (a) and the penalty term in (b).

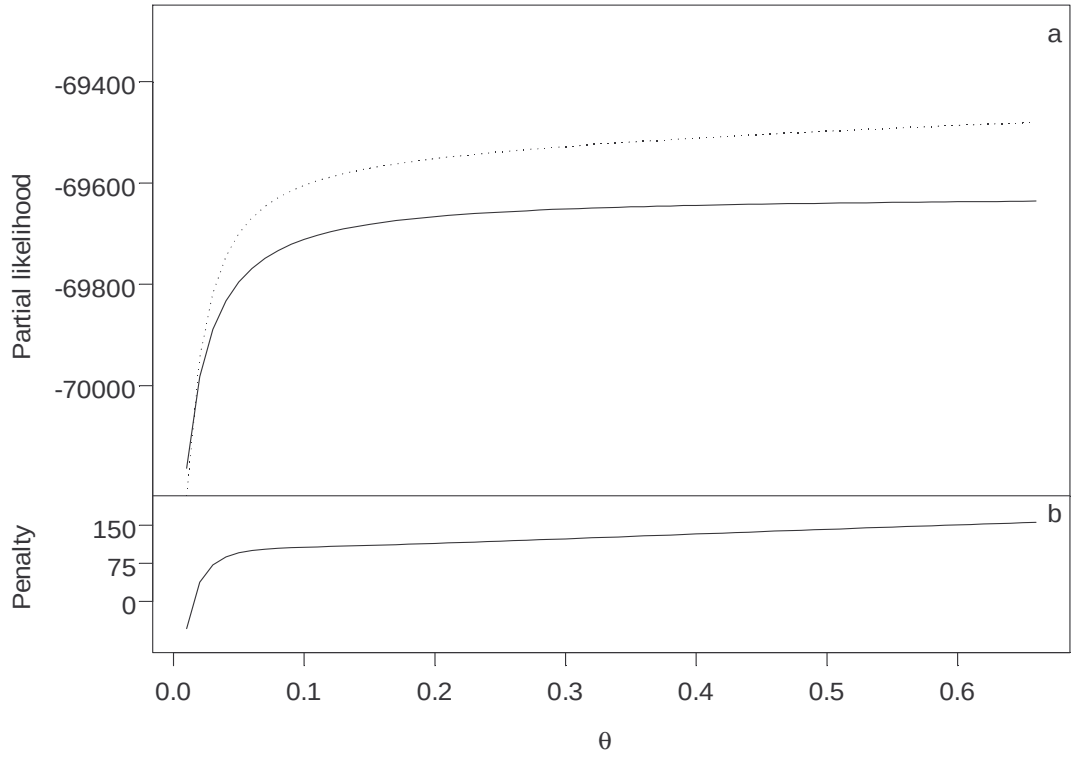


Figure 3: Profile marginal likelihood for θ with the 95% confidence interval based on the profile marginal likelihood.

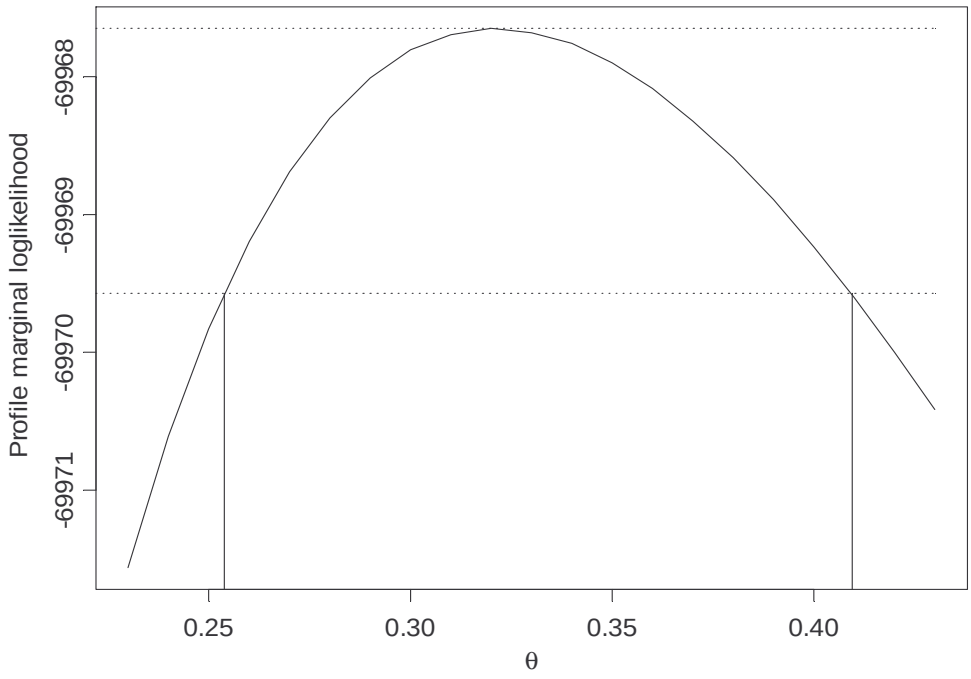


Figure 4: The relationship between the protein concentration and the relative log hazard (relative to the log hazard of the mean protein concentration). The left panel shows the density function of the milk protein concentrations in the data set.