Accelerated Failure Times Model for Arbitrarily Censored Data with Smoothed Error Distribution

Arnošt Komárek[1], Emmanuel Lesaffre[1], Joan F. Hilton[2]

[1] Catholic University Leuven, Biostatistical Centre, Kapucijnenvoer 35, B–3000, Leuven, Belgium

[2] University of California San Francisco, Dept. of Epidemiology and Biostatistics, 500 Parnassus Avenue, CA 94143–0560, San Francisco, California

Corresponding author: Arnošt Komárek

E-mail: arnost.komarek@med.kuleuven.ac.be

Tel.: +32 - 16 - 336886

Fax: +32 - 16 - 336900

In this article a procedure is developed to estimate model parameters in an accelerated failure times model (AFT) for which the error distribution is not specified. Our approach is based on a smoothing methodology and employs the technique of P-splines (B-splines with penalties) of Eilers and Marx (1996) to express a density of the error distribution. However, the error density expressed as a B-spline has a finite support, the property that might be undesirable. For this and other reasons we have replaced B-splines by normal densities whose appear to be their limits (when the degree of B-spline goes to infinity). The spline coefficients as well as the regression parameters are estimated via a penalized maximum likelihood method. The procedure allows not only for left or right censored data but also for interval censored data as is often the case in the dental or AIDS research. The approach is applied here to problems from these two areas.

KEY WORDS: Linear regression; Penalized maximum likelihood; Spline; Survival analysis.

1 INTRODUCTION

There can be little doubt that the accelerated failure times model (AFT) has a respected role in survival analysis today even though it is used far less broadly than Cox's proportional hazards model (Cox 1972). Whereas Cox's model relates the hazard function with covariates, the AFT model postulates a direct relationship between the time to event and covariates. It specifies that the effect of a vector of fixed covariates $\boldsymbol{x}$ acts multiplicatively on the time to event T, or additively on $Y = \log(T)$ as

$$\log(T) = \alpha + \boldsymbol{\beta}'\boldsymbol{x} + \sigma\varepsilon, \tag{1}$$

where α and $\boldsymbol{\beta}$ are regression parameters, σ is a scale parameter and ε is the random error with density $f(e)$. Expression (1) is simply a linear model on the log scale of the time. However, unlike the area of uncensored data where the normal distribution is the most used error distribution there is no gold standard distribution in the analysis of censored data. Moreover, in survival analysis non- or semi-parametric procedures for estimating the regression parameters are generally preferred.

In the past primarily two non-parametric methods for estimating the regression parameters have been examined in conjunction with the AFT model. The first one is based on the generalization of the least squared method to censored data first proposed by Miller (1976) and in an alternative way by Buckley and James (1979) whose names are usually used for this approach. Lai and Ying (1991) suggested a mild modification of the Buckley–James estimator and derived its asymptotic properties. The second approach is based on censored data rank tests which appeared in different forms in Prentice (1978), Gill (1980) and Louis (1981). Tsiatis (1990) extended it in a multiple regression context. Further, Ritov (1990) has pointed out an asymptotic equivalence of the Buckley–James method and the rank estimators. The asymptotic properties of the rank estimators were presented, in the greatest generality, by Ying (1993). In the case of the rank based estimators of the AFT model, the numerical aspect is a major problem. Recently, Jin et al. (2003) suggested an algorithm to compute the rank based estimates using a linear programming technique.

Both nonparametric approaches may be problematic in some practical situations, however, especially with interval censored data. Although the Buckley–James method can be rather simply extended to the area of interval censoring, this method may fail to converge or may oscillate among several solutions. On the other hand, only with considerable difficulties, the rank based estimators can be extended to handle interval censored data. We suggest an approach which is semi–parametric in its nature: $f(e)$ is estimated with a smoothing technique and regression parameters are estimated with a likelihood based method. Finally, the method can handle interval censored data in a straightforward manner.

The second section of the paper gives a brief overview of B-splines which serves as a motivation for our method. In Section 3, the basis of our approach to the approximation of the error distribution in the AFT model and the method of estimation are explained. The inference based on penalized models is described in the fourth section. In Section 5, a simulation study is presented which shows the performance of our method. Illustrations of the use of the proposed method on real data from dental and AIDS research are given in the sixth section. The paper is finalized by a discussion in the seventh section.

2 B-SPLINES AND P-SPLINES

Our approach is motivated by the approximation of the density of the error distribution by a mixture of B-splines. A brief overview of B-splines is given below followed by the concept of P-splines. More details on B-splines can be found in de Boor (1978) and Dierckx (1993), P-splines are described in detail by Eilers and Marx (1996).

Here we investigate the use of B-splines to approximate the density of the error distribution by a spline function $h(e)$. Let the domain of the spline function (e_{min}, e_{max}) be divided into $g' + 1$ equal intervals by $g' + 2$ knots: $e_{min} = \mu_0 < \cdots < \mu_{g'+1} = e_{max}$. Let $\mu_{-k} = \cdots = \mu_{-1} = e_{min}$, $\mu_{g'+2} = \cdots = \mu_{g'+k+1} = e_{max}$ be $2k$ additional boundary knots. A basis B-spline of degree k with domain (μ_j, μ_{j+k+1}) will be denoted $B_{j,k+1}$. It consists of $k + 1$ polynomial pieces of the same degree k connected at its inner knots $\mu_{j+1}, \ldots, \mu_{j+k}$ in a specific way. Namely, $B_{j,k+1}$ is positive on (μ_j, μ_{j+k+1}) and zero elsewhere. Further, its derivatives up to order $k - 1$ are continuous on $\mathbb{R}$. Except at the boundaries, each B-spline overlaps with $2k$ polynomial pieces of its neighbors. The spline function h is then expressed as

$$h(e|\boldsymbol{c}) = \sum_{j=-k}^{g'} c_j B_{j,k+1}(e),$$

where $c_{-k}, \ldots, c_{g'}$ are B-splines coefficients that need to be estimated from the data. The B-splines have the property $\sum_{j=-k}^{g'} B_{j,k+1}(e) = 1$ on (e_{min}, e_{max}). Dierckx (1993) gives simple formulas for derivatives and integrals of B-splines. All these facts can be successfully exploited when using the spline function h as the approximation for the density of the error distribution in the AFT model. Constraints to the B-spline coefficients such that $c_j > 0$ and $\sum_j c_j = (e_{max} - e_{min})^{-1}$ ensure that the resulting spline function is a density. The easiness of computation of integrals allows for the computation of the likelihood contributions even for arbitrarily censored observations.

Choosing the optimal number and position of knots is generally a complex task in the area of spline smoothing. Too many knots lead to overfitting of the data, too few knots lead to underfitting. O'Sullivan (1986, 1988) proposed to take a relatively large number of knots and to restrict the flexibility of the fitted curve by putting a penalty on the second derivative. Eilers and Marx (1996) employed this approach in the context of B-splines with a penalty term which is based on squared finite differences of higher order (in their examples, Eilers and Marx 1996 use the second order) of the spline coefficients c_i. Basically, the penalized log-likelihood is maximized instead of the ordinary log-likelihood function when estimating the parameters. Eilers and Marx (1996) suggest to take a large number of equidistant fixed knots. The term P-splines (penalized splines) was then introduced by them to indicate that the spline coefficients are estimated via maximization of the penalized log-likelihood. In the following, we will still use the term B-spline since we exploit them to describe generally the motivation for specification of our model where B-splines will be replaced by their limits. The method of estimation of spline coefficients (penalized maximum likelihood) comes consequently into consideration.

3 PENALIZED GAUSSIAN MIXTURE METHOD

3.1 Gaussian mixture as an error distribution in the AFT model

B-splines have a finite support. However, most continuous survival distributions are, on a log scale, thought of as having a support of the real line. While in practice this might not constitute any difficulty, in theory it might be more comfortable to approximate an error density having an infinite support. Further, it might be quite difficult in some settings to find a proper range of an error density with finite support since the error distribution is seen from the data only via unknown regression coefficients. However, one can easily find that the basis B-spline $B_{j,k+1}$ of degree k is proportional to the density of a sum of $k+1$ independent uniformly distributed random variables. Considering a standardized version of the basis B-spline (i.e., the basis B-spline with knots chosen such that it corresponds to a zero mean, unit variance density) denoted B_{k+1} it can be proved using the central limit theorem that B_{k+1} converges uniformly, as $k \to \infty$, on $\mathbb{R}$ to a Gaussian density (see Unser et al. 1992 for details). Moreover, the convergence is rather fast. Indeed, the standardized cubic basis B-spline B_4 is already quite close to the Gaussian density as shown in Figure 1.

< Figure 1 about here.>

This reasoning led us to replace the B-splines in the spline function h by Gaussian densities resulting in the following approximation to the density of the error term of the AFT model (1):

$$f(e|\boldsymbol{c}) = \sum_{j=1}^{g} c_j \varphi_{\mu_j,\sigma_0^2}(e), \tag{2}$$

where $\varphi_{\mu_j,\sigma_0^2}(e)$ is the Gaussian density with mean μ_j and variance σ_0^2 having now the real line as its support. When used in this context, we will call the basis functions $\varphi_{\mu_j,\sigma_0^2}$ basis Gaussian densities, or briefly BG-densities. The values $\mu_1 < \cdots < \mu_g$ retain their meaning of equidistant knots, chosen beforehand. The basis standard deviation σ_0 corresponds more or less to the choice of a degree of the B-spline and is also fixed. A set of mixture coefficients $\boldsymbol{c}$ such that

$$\sum_{j=1}^{g} c_j = 1, \qquad c_j > 0, \quad j = 1, \dots, g,$$

ensures that $f(e|\boldsymbol{c})$ is a density function.

To avoid constrained maximization one can use an alternative parametrization using coefficients $a_j, \quad j = 1, \dots, g$, i.e.

$$c_j(\boldsymbol{a}) = \frac{\exp(a_j)}{\sum_{l=1}^{g} \exp(a_l)}, \qquad j = 1, \dots, g,$$

with one of the a_j's fixed to a particular value, say $a_g = 0$. To render the intercept α and the scale σ identifiable, we need to impose the following constraints:

$$\mathrm{E}(\varepsilon|\boldsymbol{a}) = \sum_{j=1}^{g} c_j(\boldsymbol{a})\mu_j = 0, \qquad \mathrm{var}(\varepsilon|\boldsymbol{a}) = \sum_{j=1}^{g} c_j(\boldsymbol{a})(\mu_j^2 + \sigma_0^2) = 1. \tag{3}$$

It is easily seen that the basis standard deviation σ_0 must be smaller than one to be able to satisfy the variance constraint. The two equality constraints (3) can be avoided if two coefficients, say, a_{g-2} and a_{g-1}, are expressed as a function of the remaining coefficients, denoted together as a vector $\boldsymbol{d} = (a_1, \dots, a_{g-3})'$:

$$a_k(\boldsymbol{d}) = \log\Big\{\omega_{g,k} + \sum_{j=1}^{g-3} \omega_{j,k} \exp(a_j)\Big\}, \qquad k = g-2,\ g-1, \tag{4}$$

with

$$\begin{aligned}
\omega_{j,g-2} &= -\frac{\mu_j - \mu_{g-1}}{\mu_{g-2} - \mu_{g-1}} \cdot \frac{1 - \sigma_0^2 + \mu_{g-1}\mu_j}{1 - \sigma_0^2 + \mu_{g-1}\mu_{g-2}}, \\
\omega_{j,g-1} &= -\omega_{j,g-2} \cdot \frac{\mu_{g-2}}{\mu_{g-1}} - \frac{\mu_j}{\mu_{g-1}}, \qquad\qquad j = 1, \dots, g-3,\ g.
\end{aligned}$$

The density of the error distribution, a mixture of BG-densities will be further denoted as $f(e|\boldsymbol{d}) = \sum_{j=1}^{g} c_j(\boldsymbol{d})\varphi_{\mu_j,\sigma_0^2}(e)$ rather than $f(e|\boldsymbol{c})$.

With respect to the choice of the knots and the basis standard deviation σ_0 we have the following recommendation. According to the philosophy of penalized models, we use a set of a higher number of equidistant knots. The range of knots from -6 to 6 is broad enough even for distributions with heavy tails such as the extreme value distribution. To approximate a standardized continuous density with satisfactory precision, a distance of 0.3 between two consecutive knots is small enough. Moreover, a choice of $\sigma_0 = 2/3\,(\mu_{j+1} - \mu_j)$ corresponds with cubic basis B-splines such that each BG-density overlaps practically with its 6 neighbors as the cubic basis B-spline does (note that a normal density is practically zero outside $\mu \pm 3\sigma_0$ and with choice of σ_0 as above BG-density is practically zero outside (μ_{j-2}, μ_{j+2})).

All parameters in the model (transformed mixture coefficients $\boldsymbol{d}$; regression parameters α, $\boldsymbol{\beta}$; and log-scale $\log(\sigma)$) are estimated by the mean of a penalized maximum likelihood method. We construct a penalized log-likelihood function which consists of an ordinary log-likelihood and a difference penalty for the transformed spline coefficients. The penalized log-likelihood is subsequently maximized to obtain the estimates. We call our approach the *penalized Gaussian mixture method* (PGM-method).

3.2 Penalized maximum-likelihood

3.2.1 Penalized log-likelihood

Let $\boldsymbol{\theta}$ be a vector of all unknown parameters to be estimated, i.e. $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}', \log(\sigma),\, a_1, \ldots, a_{g-3})'$. Let $\ell_i(\boldsymbol{\theta}) = \ell_i(y_i|\boldsymbol{\theta})$, $i = 1, \ldots, n$ denote the ordinary log-likelihood contribution of the i-th observation based on model (1) with error density (2), with $\ell(\boldsymbol{\theta}) = \ell(\boldsymbol{y}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\theta})$. To form the penalized log-likelihood function, we subtract a penalty term $q\big(\boldsymbol{a}(\boldsymbol{d}); \lambda\big)$ for the transformed mixture coefficients $\boldsymbol{a}(\boldsymbol{d})$ from the ordinary log-likelihood

$$\ell_P(\boldsymbol{\theta}; \lambda) = \ell_P(\boldsymbol{y}|\boldsymbol{\theta}; \lambda) = \ell(\boldsymbol{\theta}) - q\big(\boldsymbol{a}(\boldsymbol{d}); \lambda\big), \qquad (5)$$

where λ is a fixed tuning parameter to control continuously the smoothness of the fitted error distribution. Suppose for the moment that its reasonable value is known. Eilers and Marx (1996) proposed to base the penalty on squared (higher-order) finite differences of the coefficients of adjacent B-splines. We base our penalty on squared finite differences of order m of the transformed coefficients of adjacent BG-densities:

$$
\begin{aligned}
q\big(\boldsymbol{a}(\boldsymbol{d});\lambda\big) &= \frac{\lambda}{2}\sum_{j=m+1}^{g}\big\{\Delta^m a_j(\boldsymbol{d})\big\}^2 \\
&= \frac{\lambda}{2}\boldsymbol{a}(\boldsymbol{d})'\boldsymbol{D}_m'\boldsymbol{D}_m\,\boldsymbol{a}(\boldsymbol{d}),
\end{aligned}
\qquad (6)
$$

where $\Delta^1 a_j = a_j - a_{j-1}$, $\Delta^m a_j = \Delta^{m-1}a_j - \Delta^{m-1}a_{j-1}$, $m = 1,\dots$ and $\boldsymbol{D}_m$ is a $(g-m)\times g$ difference operator matrix. Eilers and Marx (1996) use $m=2$ in their examples and according to our experience, $m=2$ or $m=3$ is sufficient to get smooth estimates of the density. The choice $m=3$ has another interesting justification as explained in the following paragraph and that is why we prefer it in our context.

In practice, we maximize the penalized log-likelihood first as a function of an extended parameter vector $(\alpha,\boldsymbol{\beta}',\log(\sigma),\ a_1,\dots,a_{g-3},a_{g-2},a_{g-1})'$ under the constraints (3) using the sequential quadratic programming algorithm of Han (1977) to avoid negative values in the logarithmic expression (4). After convergence is reached, we perform additional Newton–Raphson steps for the penalized log-likelihood being a function of the parameter $\boldsymbol{\theta}$ to be able to draw the inference as described in the following sections.

The estimation procedure has been implemented as a function in R environment

(`http://www.r-project.org`)

and is available upon request from the first author.

3.2.2 Remarks on the penalty function

It is useful to explain why we penalize the transformed mixture coefficients $\boldsymbol{a}$ instead of the original ones $\boldsymbol{c}$. First, with our penalty, higher differences between adjacent transformed coefficients $\boldsymbol{a}$ correspond to the area where original coefficients $\boldsymbol{c}$ are close to zero, i.e. to the area where there are not too many datapoints,

$$\begin{array}{llll} \text{e.g., for} & \breve{\boldsymbol{c}} = (0.001, 0.002, 0.001, 0.996)', & & \tilde{\boldsymbol{c}} = (0.201, 0.202, 0.201, 0.396)' \\ \text{we have} & \breve{\boldsymbol{a}} = (-6.904, -6.211, -6.904, 0)', & & \tilde{\boldsymbol{a}} = (-0.678, -0.673, -0.678, 0)' \end{array}$$

$$\begin{array}{llcl} \text{and} & (\Delta^2 \breve{a}_3)^2 = & 1.92 \gg 0.000099 & = (\Delta^2 \tilde{a}_3)^2 \\ \text{while} & (\Delta^2 \breve{c}_3)^2 = & 0.000004 & = (\Delta^2 \tilde{c}_3)^2. \end{array}$$

Indeed, in the areas with a sufficient amount of data, the estimated shape of the error distribution is more driven by the data itself and in the data-poor areas, the shape of the fitted error distribution is inter– or extrapolated from the data-rich areas as flexible as allowed by the penalty term. We also tried the penalty term based directly on the original mixture coefficients $\boldsymbol{c}$ however with less satisfactory results.

Second, the penalty of the third order ($m = 3$) based on transformed mixture coefficients $\boldsymbol{a}$ has the following interesting property. Assume that for fixed N we have a set of knots $-N$, $-N + 1/N$, $\ldots$, $-1/N$, 0, $1/N$, $N - 1/N$, N. Suppose, with a given set of knots, we maximize the penalized log-likelihood (5) for $\lambda \to \infty$. This is equivalent (in the limit) to just minimizing the penalty term (6) under the constraints (3). Let, for fixed N, f_N be the fitted error density arising from above mentioned optimization problem. It is possible to show (see the Appendix) that $\lim_{N \to \infty} f_N(e) = \varphi_{0,1}(e)$, the standard normal density. In practice, a set of knots as recommended in the previous section (knots from -6 to 6 by 0.3) with the basis standard deviation of 0.2 give rise to the fitted error density practically indistinguishable from the normal density when only the penalty term is minimized.

3.2.3 Selecting the smoothing parameter

In the area of density estimation, cross–validation type methods for selection of the smoothing parameter, λ, are usually used. Using the notation from the previous paragraph, the standard modified maximum likelihood cross–validation score which is attempting to be minimized is

$$\mathrm{CV}(\lambda) = -\sum_{i=1}^{n} \ell_i(\hat{\boldsymbol{\theta}}^{(-i)}),$$

where $\hat{\boldsymbol{\theta}}$ is the penalized MLE of $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}^{(-i)}$ the penalized MLE based on the sample with the ith observation removed. However, the computation and optimization of the cross–validation score is extremely computationally intensive in our case. On the other hand, O'Sullivan (1988) suggested, in a similar context, a one–step Newton-Raphson approximation combined with an approximation given by the first-order Taylor expansion. This method can be applied also here resulting in the approximate cross-validation score given by

$$\overline{\mathrm{CV}}(\lambda) = -\Bigl\{\sum_{i=1}^{n} \ell_i(\hat{\boldsymbol{\theta}}) - \mathrm{trace}\bigl(\hat{\boldsymbol{H}}^{-1}\hat{\boldsymbol{I}}\bigr)\Bigr\},$$

where $\hat{\boldsymbol{H}} = -\partial^2\ell_P(\hat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T$ and $\hat{\boldsymbol{I}} = -\partial^2\ell(\hat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T$. If we denote trace$(\boldsymbol{H}^{-1}\hat{\boldsymbol{I}})$ as df(λ) and interpret it as the *effective degrees of freedom* or *the effective dimension* of the model the minimization of the above expression is essentially the same as maximization of the Akaike's information criterion AIC$(\lambda) = \ell(\hat{\boldsymbol{\theta}}) - \mathrm{df}(\lambda)$ (Akaike 1974).

Depending on a chosen order m of the differences in the penalty, the degrees of freedom decreases in λ from $\dim(\boldsymbol{\beta}) + 2 + g - 3$ (if $\lambda = 0$) to $\dim(\boldsymbol{\beta}) + 2 + m - 3$ (if $\lambda \to \infty$ and $m \geq 3$). Indeed, if $\lambda \to \infty$ optimization of the ordinary log-likelihood part of the penalized log-likelihood is no more determinant to give the estimates of the transformed mixture coefficients $\boldsymbol{a}$. These are the minimizers of the penalty function (6) under the three constraints: $a_g = 0$ and (3). For a given order of the penalty $m \geq 3$ the transformed mixture coefficients $\boldsymbol{a}$ which minimize the penalty only have to lie on a curve given by the polynomial of degree $m - 1$ which is itself determined by m coefficients. For $m \geq 3$, the three constraints ($a_g = 0$ and (3)) decrease then the dimension by 3 resulting in the expression given above.

Combining the results presented in this and previous paragraph, one obtains that, with a sufficiently dense set of knots, the fitted error density is close to the normal density when the optimal value of the tuning parameter approaches infinity. This result can be subsequently used to verify a normality of the error term.

4 INFERENCE BASED ON THE PENALIZED MLE

For $\lambda > 0$, the penalized MLE $\hat{\boldsymbol{\theta}}$ is necessarily a biased estimator. For that reason, its standard errors may not be very informative if that bias is high. However, there are two possibilities how to draw the inference based on penalized MLE.

4.1 Bayesian technique

Wahba (1983) describes a Bayesian technique for generating confidence bands around the cross-validated smoothing spline. O'Sullivan (1988) used this technique in the penalized ML framework and his approach can be adopted also here. Basically, the penalized log-likelihood ℓ_P is viewed as a posterior log-density for the parameter $\boldsymbol{\theta}$ and the penalty term as a prior negative log-density of that parameter. Then, the second order Taylor expansion of the posterior log-density around its mode $\hat{\boldsymbol{\theta}}$ leads to

$$\ell_P(\boldsymbol{\theta}) \approx \ell_P(\hat{\boldsymbol{\theta}}) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \hat{\boldsymbol{H}}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}).$$

Finally the Gaussian approximation gives posterior normal distribution for $\boldsymbol{\theta}$ with covariance matrix

$$\widehat{\mathrm{var}}_B(\hat{\boldsymbol{\theta}}) = \hat{\boldsymbol{H}}^{-1}. \tag{7}$$

We call this estimate of the variance of the penalized MLE $\hat{\boldsymbol{\theta}}$ the "Bayesian variance estimate".

4.2 Asymptotical inference

More formal inference is possible under the following assumptions. First of all, we assume independent noninformative censoring (see Kalbfleisch and Prentice 2002). Further, we require that the number and position of knots are fixed as the sample size n increases and the same is true for the basis standard deviation σ_0. Let $\boldsymbol{\theta}_T$ be the true parameter value of $\boldsymbol{\theta}$, assuming at this moment that it exists. To be able to get asymptotically unbiased estimates we have to either keep the value of the smoothing parameter λ constant as $n \to \infty$ or let it increase at a rate lower than n (i.e., $\lambda = \lambda_n$ and $\lim_{n\to\infty} \lambda_n/n = 0$). Then the penalty part of the penalized log-likelihood reduces its importance relative to the log-likelihood part as $n \to \infty$. In combination with standard maximum likelihood arguments, we get that for arbitrary $\varepsilon > 0$ the penalized MLE $\hat{\boldsymbol{\theta}}$ satisfies $\mathrm{P}_{\boldsymbol{\theta}_T}\big(|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_T| < \varepsilon\big) \to 1$. Using the same arguments as in Gray (1992), one can further show that $\sqrt{n}\big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_T\big)$ is asymptotically normal with mean $\mathbf{0}$ and a covariance matrix $\lim_{n\to\infty}(n\,\boldsymbol{W})$ where the matrix $\boldsymbol{W}$ can be consistently estimated by

$$\widehat{\mathrm{var}}_A(\hat{\boldsymbol{\theta}}) = \hat{\boldsymbol{H}}^{-1}\hat{\boldsymbol{I}}\,\hat{\boldsymbol{H}}^{-1}, \tag{8}$$

which we call the "asymptotical variance estimate". As pointed out by Gray (1992), the asymptotic distribution of $\hat{\boldsymbol{\theta}}$ remains the same if the smoothing parameters λ_n are replaced by estimates satisfying $\hat{\lambda}_n/\lambda_n \xrightarrow{\mathrm{P}} 1$.

4.3 The Bayesian versus the asymptotic variance estimate

In various applications, the Bayesian variance estimate (7) has been shown to be useful. Wahba (1983) showed it had good frequentist coverage properties for pointwise intervals for the smoothing spline curve fit to noisy data. Verweij and Van Houwelingen (1994) used it in the context of penalized likelihood in Cox regression; they call the square roots of its diagonal elements "pseudo-standard errors". Joly et al. (1998) exploited this technique to get confidence bands of the hazard function smoothed using M-splines. On the other hand, for the asymptotic variance estimate (8) there is no guarantee that for finite samples its middle part $\hat{\boldsymbol{I}}$ is a positive semidefinite. Based on our experience, this problem is not rare. Finally, according to our simulations (results not shown), the Bayesian variance estimate (7) has, for regression parameters with small samples, better frequentist coverage properties than the asymptotic estimate (8).

4.4 Remarks

We have assumed in this section that the true parameter vector $\boldsymbol{\theta}_T$ exists. This does not have to be true. Especially, true $\boldsymbol{a}$ coefficients do not have to generally exist. This happens if the true error distribution is not a mixture of BG-densities determined by the choice of knots and the standard deviation σ_0. However, if the distance between the two consecutive knots is small enough we argue that the resulting Gaussian mixture can approximate every continuous distribution sufficiently well such that the assumption on the existence of the true parameter vector $\boldsymbol{\theta}_T$ is not restrictive at all.

5 SIMULATION STUDY

To see how the proposed method performs, we carried out a simulation study. The data were generated according to the model

$$\log(T) = 1.6 - 0.8 \cdot z_1 + 0.4 \cdot z_2 + 1.4 \cdot \varepsilon,$$

where the covariate z_1 was binary taking a value of 1 with probability 0.4, the covariate z_2 was generated according to the extreme value distribution of a minimum with a location 8.5 and a scale 1. The model attempts to mimic an AFT model used for WIHS dataset presented in the next section with z_1 playing the role of the covariate *dental* and z_2 being similarly distributed as $\log_2(1 + \text{CD4 count})$ in that dataset. Time to the event T corresponds to months. The error term ε was generated from a standard normal distribution $\mathrm{N}(0, 1)$, a standardized extreme value distribution and from a mixture of two normal distributions $0.4\ \mathrm{N}(-1.4, 0.8^2) + 0.6\ \mathrm{N}(0.93, 0.8^2)$. Samples of sizes 100, 300 and 600 were generated. Each simulation involved 100 replications. The censoring was created by simulating consecutive 'visit times' for each subject in the dataset. Times of the first 'visits' were drawn from $\mathrm{N}(7,\ 1)$ distribution. Distances between each consecutive 'visit' time came from $\mathrm{N}(6,\ 0.5^2)$. This approach reflexes the idea that the subject is seen for the first time about 7 months from the onset of the study and then approximately every 6th month. At each visit a prespecified ratio (between 0.4% and 0.7% for light censoring and between 4.0% and 5.0% for heavy censoring) of subjects was withdrawn from the study creating right censored observations provided that the true event time T was higher than a time of the visit at which the subject was withdrawn. For each 'true' dataset, four 'observed' datasets were created (according to the amount of drop-out and the way we viewed the data): light RC dataset where about 20% of observations were right censored and the rest exactly observed (thus ignoring the 'visit' times), light RC + IC dataset where exact observations were replaced by intervals given by the 'visit' times between which the 'true' response time fell. By increasing a proportion of subjects who dropped out at each visit, we created further heavy RC and heavy RC + IC datasets containing each about 60% right censored observations.

For comparison, estimates for each dataset were computed using our smoothed procedure and using two parametric models: AFT model with a correctly specified error distribution (normal, extreme value or mixture of normals, respectively) and AFT model with a normal error distribution. For the smoothing procedure, the third order penalty, equidistant knots with a distance of 0.3 between the two consecutive ones, and the basis standard deviation of 0.2 were used. Selected results, corresponding to the least favorable settings (the censoring patterns involving interval censoring with extreme value and mixture of normals error distributions) are presented in Table 1. Plots of average fitted error distributions for the selected simulation patterns together with their 95% pointwise simulation-based confidence bands are shown on Figure 2.

`< Table 1 about here.>`

`< Figure 2 about here.>`

Based on the results of the simulation, with light censoring even the smallest sample gave better estimates of regression parameters when used with smoothed error distribution than when used with the incorrect normal error distribution. Moreover, the MSE's of the estimates obtained by our smoothing procedure are approaching the MSE's of the model with correctly specified error distribution. On the other hand, with heavy censoring smoothed estimates based on the samples including interval censored observations were slightly worse than the estimates from the model with incorrectly specified error distribution in the case of a normal mixture as the true error distribution. When the extreme value distribution was the truth, estimates of regression parameters based on the smoothed procedure were only slightly worse and only in the presence of interval censored observations than the estimates obtained by assuming the normal error. As the fitted error distribution concerns, these were reproduced rather satisfactory, also with heavy censoring.

6 REAL DATA EXAMPLES

To illustrate applications of our method to real data, we present two examples, one from dentistry and one from AIDS research. We chose these two areas since they both often involve interval censored survival data, implying that the traditional model Cox model given by Cox (1972) cannot be applied. However, both applications could also be analyzed using a fully parametric AFT model.

To obtain the results shown below, we used a sequence of 41 equidistant knots from -6 to 6 with a distance of 0.3 between each pair. The basis was 0.2 and the third order difference was used in the penalty. Different models were compared using Akaike's information criterion and claims concerning the significance of the parameters were based on Wald's tests. All reported standard errors are based on the Bayesian variance estimate (7).

6.1 Signal Tandmobiel Study

In the Signal Tandmobiel study, oral health data from 4 468 Flemish schoolchildren born in 1989 were collected by a team of 16 dentists, operating from a dental bus, who visited 179 schools in Flanders (the Northern region of Belgium) every year (1996 – 2001). Moreover, data from a questionnaire (given to the parents) on dietary and oral hygiene habits were obtained. More information on the study can be found in Vanobbergen et al. (2000). The data used in this paper can be obtained upon request from the corresponding author.

Leroy et al. (2003) tested the hypothesis whether caries found on the primary molars have an impact on the emergence times of permanent molars. They found that caries found on primary teeth decrease the emergence time of their permanent successors. Here, we show the analysis of the emergence time of another tooth: the permanent left mandibular canine (tooth 33 in European dental notation). This tooth emerges mainly between 8 and 11 years of age, the period covered by our study. The distribution of the emergence time is undoubtedly gender dependent. To examine a similar hypothesis as Leroy et al. (2003), we analyzed whether the emergence time of the permanent canine depends on the binary covariate, *caries*, taking the value of one if the primary predecessor (primary canine) had caries. The time that a child starts to brush its teeth could affect the outcome and so the initial brushing age was also included as a covariate. It was obtained from the questionnaire in a categorized form as ≤ 1, $(1, 2]$, $(2, 3]$, $(3, 4]$, $(4, 5]$ and > 5 years of age (covariate *stbrush*). As a response, we use the interval censored emergence time decreased by 5, the age below which is biologically impossible to see the emergence of tooth 33. Values of all considered covariates were available for 3 769 children.

< Table 2 about here.>

< Table 3 about here.>

The values of AIC for fitted models are shown in Table 2. We see that inclusion of the main effect of *caries* improved slightly the AIC for a model with *gender* as the only covariate. However, Wald tests for the significance of these terms are not significant (see Table 3 for the results). On the other hand, the inclusion of the initial brushing age does not improve the Model (1) of Table 2 in any way ($p = 0.25$ in Model (4)).

< Figure 3 about here.>

The lower part of Figure 3 shows a good agreement between the predicted cumulative distribution functions and the non–parametric Turnbull estimates (Turnbull 1976). Indeed, the difference between the CDFs for permanent teeth with decayed or sound primary predecessors is rather low. The plots of the fitted error densities (upper part of Figure 3) suggest that AFT models with normal or logistic error distribution would also be appropriate here.

In conclusion, after adjusting the emergence time of the studied permanent tooth for the biological difference between girls and boys, there is almost no effect of caries in the primary predecessor and no effect of initial brushing age on the emergence of the permanent successor.

6.2 The Women's Interagency HIV Study

To illustrate an application of our method to AIDS research, we analyzed a subsample of the HIV-seropositive women participating in the Women's Interagency HIV Study (WIHS). The total study population (over 3000 participants) was enrolled between October 1994 and November 1995 through six clinical consortia at 23 sites throughout the United States. More information on the setup of the study can be found in Barkan et al. (1998). Our subsample consisted of the 224 AIDS-free women who participated in the dental sub-study and for whom the viral load, CD4 count and the value of the dental marker (described below) were available at the baseline visit.

For HIV positive people, it is of interest to describe the distribution of the time to the onset of an AIDS related illness based on some measured quantities. Classically used predictors include the number of copies of the HIV RNA virus and the count of CD4 T-cells per *ml* of blood. We examined whether presence of one of the three dental markers, oral candidiasis, hairy leukoplakia and angular cheilitis, is useful, possibly together with one or both laboratory predictors, in describing the distribution of the residual time to onset of AIDS.

As a response, we used the time in months between the baseline visit, defined as the first visit at which the dental markers were collected by dental professionals, and the onset of an AIDS-related illness. Clinical AIDS diagnoses were self-reported in 73.5% of cases, presumptive or definitive in 17.5%, and indeterminate in 9%; the case definition did not depend on CD4 T-lymphocytes. For 66 cases the response was interval censored, while for 158 cases it was right censored. The average length of the interval between two examinations at which AIDS could be detected was 7 months. The average follow-up time was 41 months, the maximal follow-up time was 84 months.

The three dental markers were summarized in one binary covariate, *dental*, equal to one if at least one of the above mentioned three dental markers was present. We also analyzed functions of the viral load and the CD4 count in the AFT model (i.e., $lvload = \log_{10}(1 + \text{viral load})$ and $lcd4 = \log_2(1 + \text{CD4 count})$). All three covariates are moderately to strongly associated with one another since, as AIDS progresses, viral load increases, CD4 count falls, and oral lesions occur more frequently. In our sample, for women with $dental = 0$ and 1, respectively, the median *lvload* was 3.60 and 4.23 (Mann-Whitney *p*-value, 0.001), and the median *lcd*4 was 8.85 and 8.52 (Mann-Whitney *p*-value, 0.005). There was also moderate negative correlation of -0.46 between *lcd*4 and *lvload*. These associations have to be taken into account when interpreting the results. Summary of the fitted models is shown in Table 4.

< Table 4 about here.>

< Figure 4 about here.>

If used alone (model (1) in Table 4) the effect of *dental* on the time to onset of AIDS is statistically significant ($p = 0.018$) and the estimated time is $\exp(-0.87) \approx 0.42$ times shorter for women with $dental = 1$ than women with $dental = 0$. The survival curves for the two groups based on model (1) are shown on the upper left panel of Figure 4. The fitted smooth survivor functions are overlaid with the non–parametric Turnbull estimates (Turnbull 1976). The two estimates are quite close to each other, illustrating the semiparametric nature of our approach. However, our procedure gives smooth estimates of the survival curves and moreover enables quantification of the difference in survival between the two groups.

According to the AIC values for models (2) and (3) in Table 4, the transformed CD4 count and viral load are equally good predictors of the time to onset of AIDS. Addition of the dental marker improves the model with $lcd4$ considerably but improves the model with *lvload* only slightly.

Further, Figure 4 shows fitted error densities of the models of Table 4. The first four fitted error densities more or less coincide with the normal density. This is not surprising since the optimal tuning parameter λ for these models was equal to $224 \cdot \exp(2)$, essentially a value of infinity in this practical situation and thus implying fitted error distributions close to the normal distribution, as pointed out in paragraph 3.2.3. On the other hand, models where $lcd4$ was used in combination with other covariates gave much lower optimal tuning parameters λ, implying also less smooth error densities with three modes. The phenomenon of three modes in the fitted error density of models (5) to (7) could indicate presence of a risk-group mixture in the data or absence of another important predictor. Indeed, a factor that could play an important role is antiretroviral therapy, which might have been used by some women in our sample. However, this factor requires modelling time–dependent covariates, which cannot be done with our model.

In conclusion, the presence of oral lesions notably decreases the time to AIDS onset in this study population. Further, this dental marker improves the prediction of that time based on any of the classical indicators (CD4 count and viral load). Only a limited number of WIHS women opted to participate in the Oral Substudy, the source of the dental data. Thus they may differ in unknown ways from the overall set. Nonetheless, our findings are consistent with those of others who have evaluated oral lesions as predictors of AIDS onset and they illustrate use of our method in the area of AIDS research. Our method restricts us to analysis of baseline covariates. Although this is a very widely applicable special case, extension of the method to accommodate time-dependent covariates would allow more complex relationships between outcomes and covariates.

7 DISCUSSION

We have suggested and implemented as an R function a method useful to fit the linear regression model for censored observations while avoiding too restrictive parametric assumptions for the error distribution. Most classically, the logarithmic transformation of the response leads to a well known AFT model. However, other transformations of the response leading to its potential range covering the whole real line are also possible. A density of the error distribution is specified in a semi-parametric way as a mixture of basis Gaussian densities (Gaussian densities with given means – knots and given standard deviation). Mixture coefficients are then estimated using the penalized maximum–likelihood method. Such model specification gives enough flexibility with respect to the resulting error distribution yet remains tractable such that the data carrying censoring of several types and especially the interval censoring can be handled naturally.

The PGM-method was successfully used also by Ghidey et al. (2003) in the context of the linear mixed model. They exploited a mixture of BG-densities to approximate a density of the distribution of the random intercept and slope in the linear mixed model while assuming standard normal distribution for the random error. They did not assume censored observations, however, they showed an additional potential of this method by using a tensor product of two univariate mixtures of BG-densities to approximate a bivariate distribution. Generally, it would be interesting to join their and our model to form an AFT model with random effects whose distribution would be approximated by a mixture of BG-densities and with the error distribution approximated by another mixture of BG-densities as well. The question that remains and that needs additional research is the amount of computational problems that could be encountered with such a complex model.

Further, in some specific situations, it can be desirable to have a finite support for the density of the error distribution and keep the finite support also in the estimated model. The mixture of BG-densities as presented in this paper is then inappropriate. Then, one can use again a mixture of B-splines and otherwise unchanged methodology while keeping the computational complexity on the same level as with the mixture of BG-densities. However, we think that situations when the finite support for the error distribution in survival models is required is rather rare and that is why we did not implemented our methodology using B-splines.

In the literature, Kooperberg and Stone (1992) and Eilers and Marx (1996) considered a spline estimation of the logarithm of a density based on a set of i.i.d. observations. Kooperberg and Stone (1992) allow also for censoring. Their approach could be generally extended in the regression context. However, with their method a logarithm of an error density ($\log f(e)$) would be expressed as a spline ($s(e)$). When computing the likelihood contributions for censored observations one would have to evaluate an integral of the form $\int e^{s(e)}\,de$. That integral would generally ask for advanced numerical methods. This was avoided by our technique since all integrals needed to evaluate the likelihood are expressed as a linear combination of values of the cumulative distribution functions of the normal distribution and quantities related to the normal distribution can be computed using fast and rather precise numerical methods.

Finally, it is useful to stress that a mixture of BG-densities is different from a classical Gaussian mixture. In the latter case, the means and the standard deviation of Gaussian components also have to be estimated. In contrast with the classical mixture, we always use relatively high number of components and drive the smoothness of the resulting error distribution via a penalty term in the log-likelihood. Indeed, the price we pay is a relatively high number of parameters in the resulting model. However, we are able to maximize the (penalized) log-likelihood directly, without needing to use EM type algorithms.

ACKNOWLEDGEMENTS

This work was primarily supported by Research Grants OE/03/29 and OT/00/35, Catholic University Leuven. The first two authors acknowledge support from the Interuniversity Attraction Poles Program P5/24 – Belgian State – Federal Office for Scientific, Technical and Cultural Affairs. Funding for Dr. Hilton's research was provided by the U.S.A. National Institutes of Health (DHHS NIAID R01-AI55085 and NIDCR P01-DE07946).

Data collection for the Signal Tandmobiel project was supported by Unilever, Belgium. The Signal Tandmobiel project comprises the following partners: D. Declerck (Dental School, Catholic University Leuven), L. Martens (Dental School, University Ghent), J. Vanobbergen (Oral Health Promotion and Prevention, Flemish Dental Association), P. Bottenberg (Dental School, University Brussels), E. Lesaffre (Biostatistical Centre, Catholic University Leuven), K. Hoppenbrouwers (Youth Health Department, Catholic University Leuven; Flemish Association for Youth Health Care).

Data used in paragraph 6.2 of this manuscript were collected by the Women's Interagency HIV Study (WIHS) Collaborative Study Group and its Oral Substudy with centers (Principal Investigators) at New York City/Bronx Consortium (K. Anastos, J. A. Phelan); Brooklyn, NY (H. Minkoff); Washington DC Metropolitan Consortium (M. Young); The Connie Wofsy Study Consortium of Northern California (R. Greenblatt, D. Greenspan, J. S. Greenspan); Los Angeles County/Southern California Consortium (A. Levine, R. Mulligan, M. Navazesh); Chicago Consortium (M. Cohen, M. Alves); Data Coordinating Center (A. Muñoz). The WIHS is funded by the National Institute of Allergy and Infectious Diseases, with supplemental funding from the National Cancer Institute, the National Institute of Child Health & Human Development, the National Institute on Drug Abuse, the National Institute of Dental and Craniofacial Research, the Agency for Health Care Policy and Research, the National Center for Research Resources, and the Centers for Disease Control and Prevention. U01-AI-35004, U01-AI-31834, U01-AI-34994, U01-AI-34989, U01-HD-32632 (NICHD), U01-AI-34993, U01-AI-42590, M01-RR00079, and M01-RR00083. The WIHS Oral Substudy is funded by the National Institute of Dental and Craniofacial Research.

APPENDIX

Theorem. *Let* $\forall N \in \mathbb{N}$ $\boldsymbol{\mu}^N = \{\mu_j^N = \frac{j}{N},\quad j = -N^2, \dots, N^2\}$ *be a sequence of knots. Let* $\forall N \in \mathbb{N}$ *and* $\forall \boldsymbol{a} \in \mathbb{R}^{2N^2+1}$ *a discrete distribution on* $\boldsymbol{\mu}^N$ *be given by* $P(\mu^N = \mu_j^N | \boldsymbol{a}) = \exp(a_j)$. *Let* $\boldsymbol{a}^N$ *minimizes* $\sum_{j=-N^2+3}^{N^2} \{\Delta^3 a_j\}^2$ *under the constraints* $\sum_{j=-N^2}^{N^2} P(\mu^N = \mu_j^N | \boldsymbol{a}) = 1$, $\mathrm{E}[\mu^N | \boldsymbol{a}] = 0$ *and* $\mathrm{var}(\mu^N | \boldsymbol{a}) = 1 - \sigma_0^2$ *for* $\sigma_0 \in (0,1)$ *fixed. Let* $f_N(e) = \sum_{j=-N^2}^{N^2} P(\mu^N = \mu_j^N | \boldsymbol{a}^N)\, \varphi_{\mu_j^N, \sigma_0^2}(e),\ \ e \in \mathbb{R}$. *Then* $\forall e \in \mathbb{R}$ $\lim_{N\to\infty} f_N(e) = \varphi_{0,1}(e)$.

Sketch of the proof. It is easily seen that a_j^N is a quadratic function of knots, i.e. $a_j^N = b_0^N - b_2^N(\mu_j^N - b_1^N)^2$ with $b_1^N = 0$, $b_0^N = -\log\Big\{\sum_{j=-N^2}^{N^2} \exp\big[-b_2^N(\mu_j^N)^2\big]\Big\}$ and b_2^N being a solution to $C_N(b) = 0$, where

$$C_N(b) = \Big\{ \sum_{j=-N^2}^{N^2} (\mu_j^N)^2 \exp\big[-b\,(\mu_j^N)^2\big] \Big\} \cdot \Big\{ \sum_{j=-N^2}^{N^2} \exp\big[-b\,(\mu_j^N)^2\big] \Big\}^{-1} - (1 - \sigma_0^2).$$

The function $C_N(b)$ is continuous on $[0,\infty)$, $\frac{d}{db} C_N(b) = \Big\{\mathrm{E}\big[(\mu^N)^2 | b_2^N = b\big]\Big\}^2 - E\big[(\mu^N)^4 | b_2^N = b\big] < 0$ $\forall b \in [0,\infty)$, $C_N(0) = 1/3\,(N^2+1) - (1-\sigma_0^2) > 0\ \ \forall N \geq 2$ and $\lim_{b\to\infty} C_N(b) = -(1-\sigma_0^2) < 0$ $\forall N \in \mathbb{N}$. So that $\forall N \geq 2$ there exists exactly one root $b_2^N \in (0,\infty)$ of the equation $C_N(b) = 0$.

Let $C(b)$ be defined as

$$C(b) = \Big[\int_{-\infty}^{\infty} s^2 \exp\big(-b\,s^2\big)\,ds\Big] \cdot \Big[\int_{-\infty}^{\infty} \exp\big(-b\,s^2\big)\,ds\Big]^{-1} - (1-\sigma_0^2) = (2b)^{-1} - (1-\sigma_0^2).$$

The equation $C(b) = 0$ has a unique solution $b_2 = \big[2(1-\sigma_0^2)\big]^{-1} \in (1/2, \infty)$.

From properties of the integral $\forall b \in (0,\infty)$ $\lim_{N\to\infty} C_N(b) = C(b)$. Consequently, we can prove that $\lim_{N\to\infty} b_2^N = b_2$.

Let $G_N(\mu)$ be a cumulative distribution function of μ^N under b_2^N, i.e.

$$G_N(\mu) = \Big\{ \sum_{j=-N^2}^{\min(N\mu, N^2)} \exp\big[-b_2^N\,(\mu_j^N)^2\big] \Big\} \cdot \Big\{ \sum_{j=-N^2}^{N^2} \exp\big[-b_2^N\,(\mu_j^N)^2\big] \Big\}^{-1}$$

and $\Phi_{0,1-\sigma_0^2}(\mu)$ be a cumulative distribution function of the normal distribution $\mathrm{N}(0, 1-\sigma_0^2)$, i.e.

$$\Phi_{0,1-\sigma_0^2}(\mu) = \Big[\int_{-\infty}^{\mu} \exp\big(-b_2\,s^2\big)\,ds\Big] \cdot \Big[\int_{-\infty}^{\infty} \exp\big(-b_2\,s^2\big)\,ds\Big]^{-1}.$$

It can be now shown that $\forall \mu \in \mathbb{R}$ $\quad \lim_{N\to\infty} G_N(\mu) = \Phi_{0,1-\sigma_0^2}(\mu)$, i.e. the random variable μ^N under b_2^N converges in distribution to a $\mathrm{N}(0, 1-\sigma_0^2)$ random variable.

Finally, $\forall e \in \mathbb{R}\ \ f_N(e) = \int_{-\infty}^{\infty} \varphi_{\mu,\sigma_0^2}(e)\, dG_N(\mu)$ and $\varphi_{0,1}(e) = \int_{-\infty}^{\infty} \varphi_{\mu,\sigma_0^2}(e)\, d\Phi_{0,1-\sigma_0^2}(\mu)$. The assertion of the theorem now follows from the fact that $\forall e \in \mathbb{R}$ function $\varphi_{\mu,\sigma_0^2}(e)$ is a bounded and continuous function of μ.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control,* **AC–19,** 716–723.

Buckley, J., and James, I. (1979). Linear regression with censored data. *Biometrika,* **66,** 429–436.

Barkan, S. E., Melnick, S. L., Preston–Martin, S., Weber, K., Kalish, L. A., Miotti, P., Young, M., Greenblatt, R., Sacks, H., and Feldman, J. (1998). The Women's Interagency HIV Study. *Epidemiology,* **9,** 117–125.

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

de Boor, C. (1978). *A Practical Guide to Splines.* Springer, Berlin.

Dierckx, P. (1993). *Curve and Surface Fitting with Splines.* Clarendon, Oxford.

Eilers, P. H. C., and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science,* **11,** 89–121.

Gill, R. D. (1980). *Censoring and Stochastic Integrals.* Math Centre Tract 124. Math. Centrum, Amsterdam.

Ghidey, W., Lesaffre, E., and Eilers, P. (2003). P-spline smoothing of the random effects distribution in a linear mixed model. *Submitted to publication.*

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

Han, S. P. (1977). A globally convergent method for nonlinear programming. *Journal of Optimization Theory and Applications,* **22,** 297–309.

JIN, Z., LIN, D. Y., WEI, L. J., and YING, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika,* **90,** 341–353.

JOLY, P., COMMENGES, D., and LETENNEUR, L. (1998). A penalized likelihood approach for arbitrarily censored and truncated data: application to age–specific incidence of dementia. *Biometrics,* **54,** 185–194.

KALBFLEISCH, J. D., and PRENTICE, R. L. (2002). *The Statistical Analysis of Failure Time Data, 2nd ed.* Chichester: John Wiley & Sons.

KOOPERBERG, C., and STONE, C. J. (1992). Logspline density estimation for censored data. *Journal of Computational and Graphical Statistics,* **1,** 301–328.

KULLBACK, S., and LEIBLER, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics,* **22,** 79–86.

LAI, T. L., and YING, Z. (1991). Large sample theory of a modified Buckley-James estimator for regression analysis with censored data. *The Annals of Statistics,* **19,** 1370–1402.

LEROY R., BOGAERTS K., LESAFFRE E., and DECLERCK D. (2003). Impact of caries experience in the deciduous molars on the emergence of the successors. *European Journal Oral Sciences,* **111**, 106-110.

LOUIS, T. A. (1981). Nonparametric analysis of an accelerated failure time model. *Biometrika,* **68,** 381–390.

MILLER, R. G. (1976). Least squares regression with censored data. *Biometrika,* **63,** 449–464.

PRENTICE, R. L. (1978). Linear rank tests with right censored data. *Biometrika,* **65,** 167–179.

O'SULLIVAN, F. (1986). A statistical perspective on ill-posed inverse problem (with discussion). *Statistical Science,* **1,** 505–527.

O'SULLIVAN, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM Journal of Scientific Computing,* **9,** 363–379.

RITOV, Y. (1990). Estimation in a linear regression model with censored data. *The Annals of Statistics,* **18,** 303–328.

TSIATIS, A. A. (1990). Estimating regression parameters using linear rank tests for censored data.

The Annals of Statistics, **18,** 354–372.

TURNBULL, B. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B,* **37,** 290–295.

UNSER, M., ALDROUBI, A., and EDEN, M. (1992). On the asymptotic convergence of B-spline wavelets to Gabor functions. *IEEE Transactions on Information Theory,* **38,** 864–872.

VANOBBERGEN, J., MARTENS, L., LESAFFRE, E., and DECLERCK, D. (2000). A longitudinal intervention health promotion study in Flanders (Belgium): baseline and first year results. *European Journal of Paediatric Dentistry,* **2,** 87–96.

VERWEIJ, P. J. M., and VAN HOUWELINGEN, H. C. (1994). Penalized likelihood in Cox regression. *Statistics in Medicine,* **13,** 2427–2436.

WAHBA, G. (1983). Bayesian "confidence intervals" for the cross–validated smoothing spline. *Journal of the Royal Statistical Society, Series B,* **45,** 133–150.

YING, Z. (1993). A large sample study of rank estimation for censored regression data. *The Annals of Statistics,* **21,** 76–99.

Figure 1: Standardized basis B-splines (solid line) and their comparison to the Gaussian density (dashed line). Small bullets indicate inner knots, large bullets boundary knots.

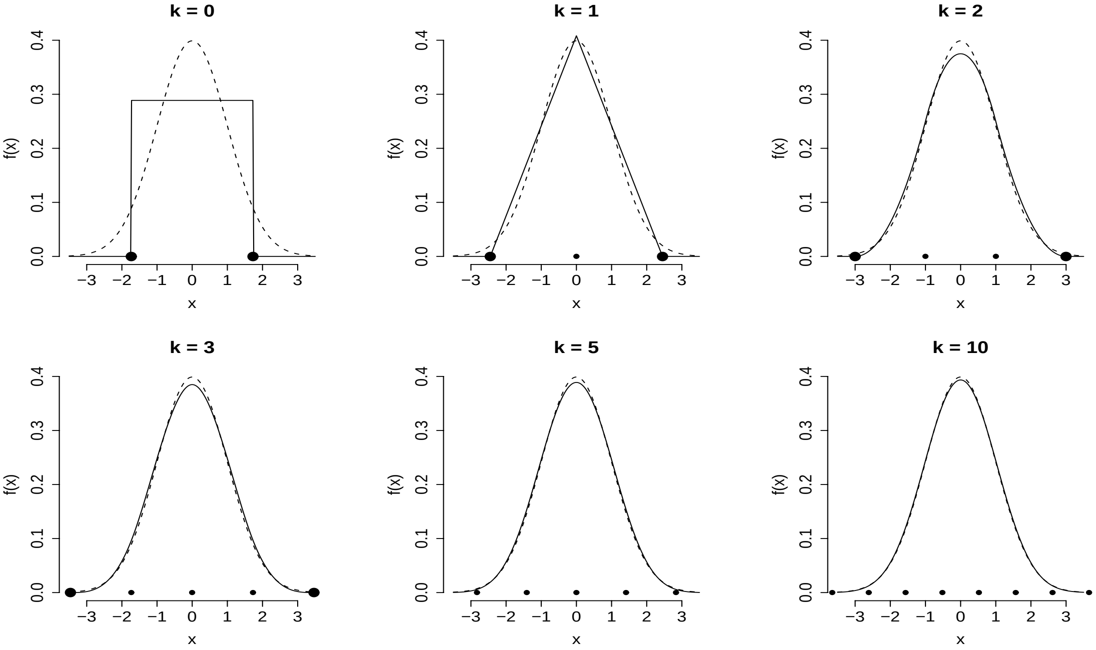

Figure 2: Simulation Study. The average of the fitted error density (solid line), 95% pointwise confidence interval (dotted line) and the true error density (dashed line) for selected simulation patterns. Extreme value as the true distribution in the upper part, normal mixture as the error distribution in the bottom part.

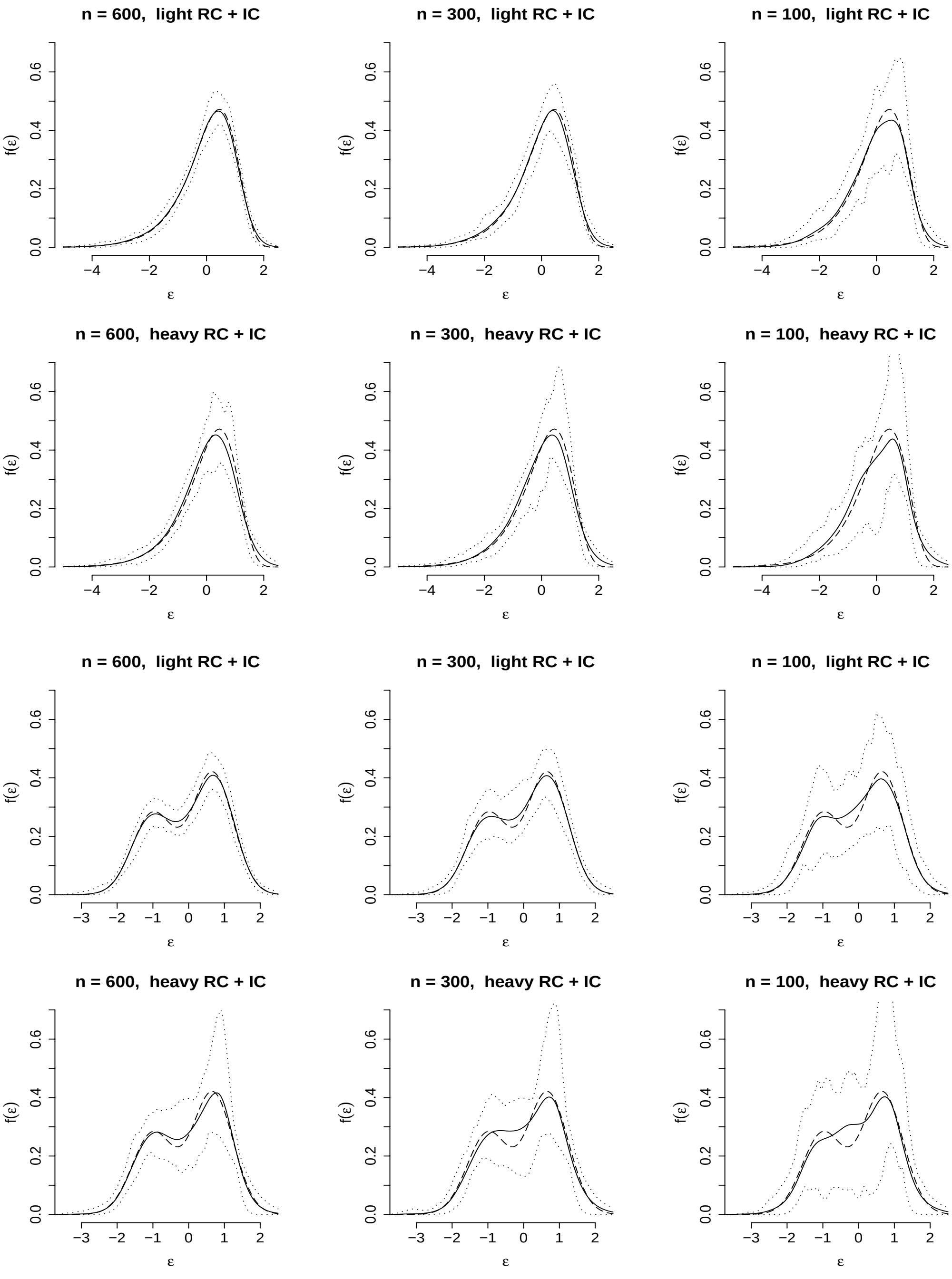

Figure 3: Signal Tandmobiel Study. Upper plots: Fitted error densities for Models (1) and (2) compared to three parametric densities: normal (dashed line), logistic (dotted line) and extreme value (dot-dashed line). Lower plots: Predicted cumulative distribution functions for the permanent left mandibular canine based on Models (1) and (2) (solid curves) compared to the non–parametric Turnbull estimates (dashed curves).

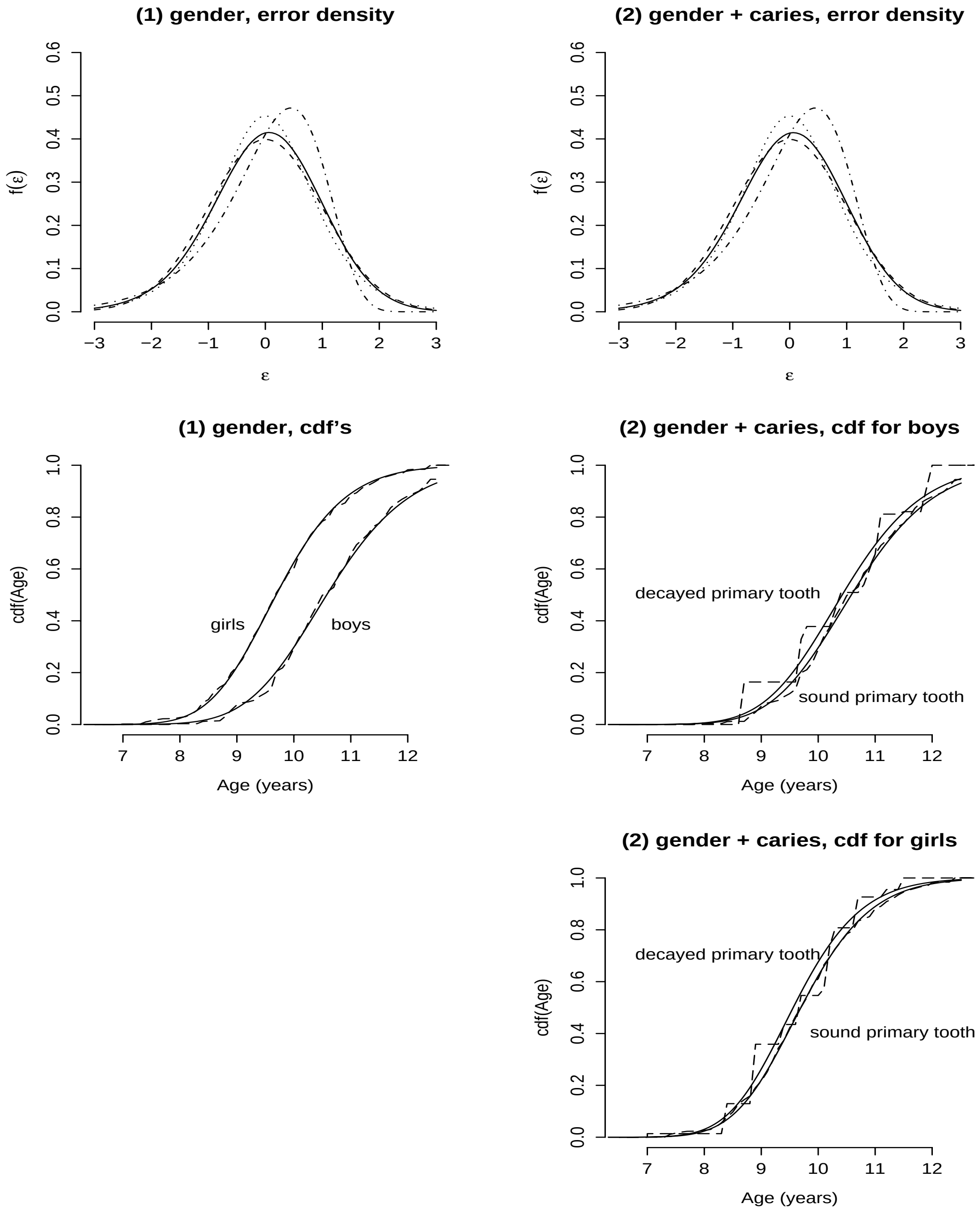

Figure 4: WIHS Data. Predicted survivor curves based on Model (1) for women with *dental* = 1 vs. women with *dental* = 0 (solid curves) compared to the non–parametric estimates of Turnbull (dashed curves) in upper right panel. Fitted error densities compared to three parametric densities: normal (dashed line), logistic (dotted line) and extreme value (dot-dashed line) in remaining panels.

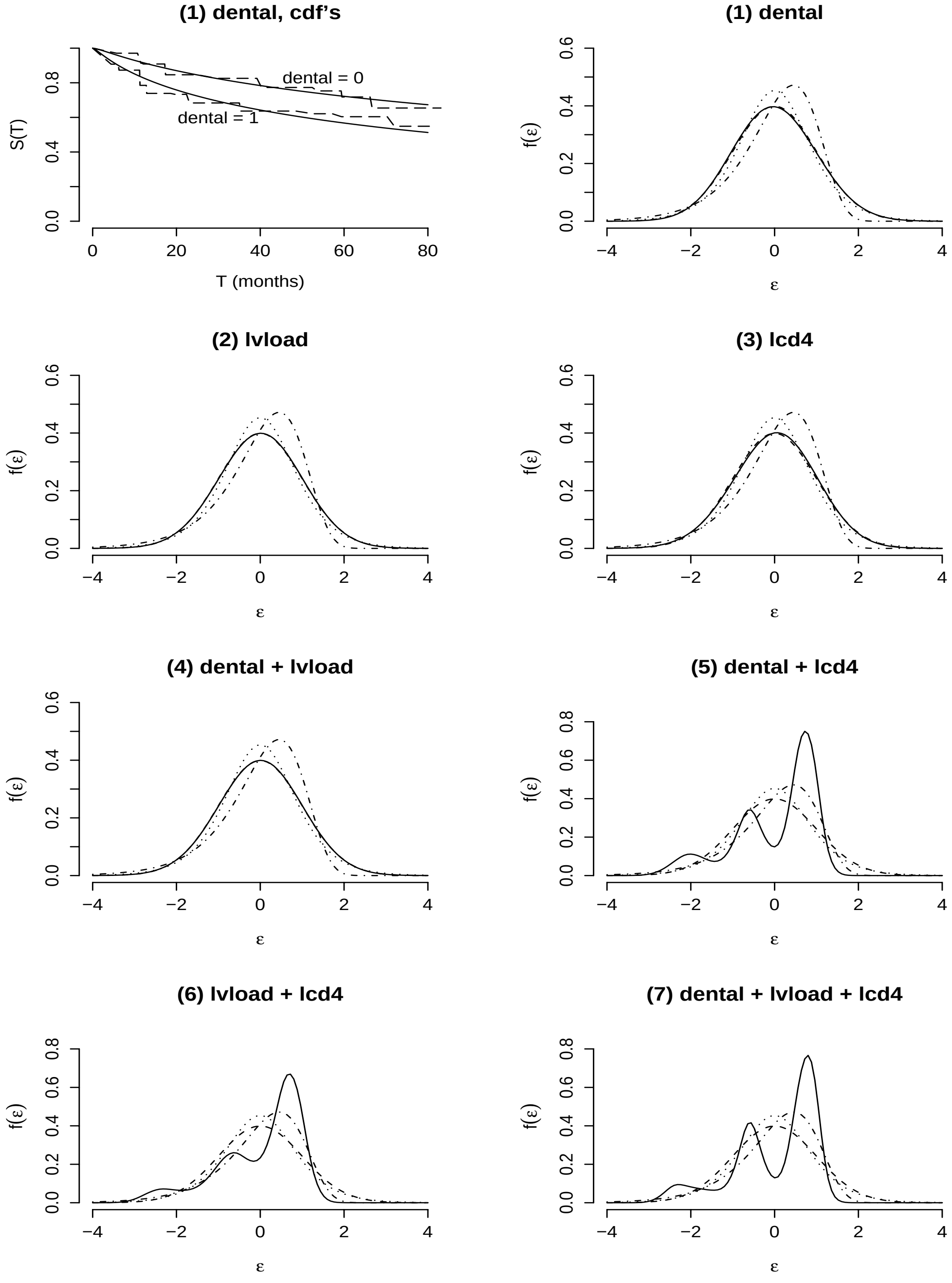

Table 1: Simulation Study.

	Assumed Error Distribution					
	Smoothed		True		Normal	
Sample size and pattern	$\hat{\beta}$ (SD)	MSE	$\hat{\beta}$ (SD)	MSE	$\hat{\beta}$ (SD)	MSE
	Binary covariate, $\beta_1 = -0.800$					
	True $\varepsilon \sim$ standardized extreme value					
600, light RC + IC	-0.795 (0.109)	0.01186	-0.786 (0.104)	0.01096	-0.793 (0.123)	0.01528
300, light RC + IC	-0.826 (0.156)	0.02501	-0.824 (0.151)	0.02327	-0.833 (0.173)	0.03099
100, light RC + IC	-0.796 (0.299)	0.08965	-0.782 (0.266)	0.07121	-0.808 (0.320)	0.10246
600, heavy RC + IC	-0.800 (0.156)	0.02423	-0.786 (0.138)	0.01918	-0.819 (0.152)	0.02336
300, heavy RC + IC	-0.855 (0.229)	0.05522	-0.856 (0.203)	0.04423	-0.880 (0.229)	0.05898
100, heavy RC + IC	-0.853 (0.469)	0.22258	-0.786 (0.368)	0.13578	-0.833 (0.420)	0.17785
	True $\varepsilon \sim 0.4\ \mathrm{N}(-1.4, 0.8^2) + 0.6\ \mathrm{N}(0.93, 0.8^2)$					
600, light RC + IC	-0.824 (0.159)	0.02585	-0.819 (0.150)	0.02290	-0.877 (0.184)	0.03992
300, light RC + IC	-0.834 (0.226)	0.05232	-0.819 (0.201)	0.04088	-0.880 (0.283)	0.08659
100, light RC + IC	-0.803 (0.411)	0.16866	-0.803 (0.323)	0.10432	-0.833 (0.466)	0.21847
600, heavy RC + IC	-0.826 (0.263)	0.07002	-0.808 (0.223)	0.04975	-0.821 (0.230)	0.05316
300, heavy RC + IC	-0.789 (0.376)	0.14128	-0.759 (0.342)	0.11891	-0.782 (0.366)	0.13459
100, heavy RC + IC	-0.752 (0.640)	0.41180	-0.776 (0.548)	0.30124	-0.779 (0.609)	0.37114
	Continuous covariate, $\beta_2 = 0.400$					
	True $\varepsilon \sim$ standardized extreme value					
600, light RC + IC	0.403 (0.041)	0.00172	0.400 (0.039)	0.00155	0.404 (0.045)	0.00201
300, light RC + IC	0.416 (0.059)	0.00379	0.412 (0.056)	0.00330	0.417 (0.067)	0.00482
100, light RC + IC	0.416 (0.101)	0.01037	0.409 (0.093)	0.00872	0.410 (0.105)	0.01119
600, heavy RC + IC	0.413 (0.061)	0.00385	0.403 (0.059)	0.00349	0.426 (0.066)	0.00504
300, heavy RC + IC	0.432 (0.084)	0.00805	0.420 (0.077)	0.00637	0.440 (0.087)	0.00923
100, heavy RC + IC	0.419 (0.164)	0.02715	0.405 (0.143)	0.02033	0.425 (0.151)	0.02348
	True $\varepsilon \sim 0.4\ \mathrm{N}(-1.4, 0.8^2) + 0.6\ \mathrm{N}(0.93, 0.8^2)$					
600, light RC + IC	0.408 (0.059)	0.00359	0.407 (0.056)	0.00317	0.424 (0.076)	0.00628
300, light RC + IC	0.408 (0.079)	0.00629	0.401 (0.071)	0.00502	0.432 (0.098)	0.01056
100, light RC + IC	0.403 (0.183)	0.03368	0.397 (0.152)	0.02309	0.417 (0.196)	0.03864
600, heavy RC + IC	0.410 (0.091)	0.00843	0.402 (0.084)	0.00699	0.401 (0.096)	0.00920
300, heavy RC + IC	0.418 (0.107)	0.01170	0.408 (0.095)	0.00900	0.405 (0.113)	0.01283
100, heavy RC + IC	0.434 (0.302)	0.09249	0.427 (0.249)	0.06288	0.418 (0.267)	0.07137

Table 2: Signal Tandmobiel Study. Akaike's information criterion, degrees of freedom and the optimal $\log(\lambda)$ for the fitted models.

Model	AIC	*df*	$\log(\lambda/n)$
(1) *gender*	$-4\,643.81$	4.7	0
(2) *gender* + *caries*	$-4\,643.80$	5.7	0
(3) *gender* + *caries* + *gender*×*caries*	$-4\,644.72$	6.7	0
(4) *gender* + *stbrush*	$-4\,645.49$	9.7	0
(5) *gender* + *stbrush* + *gender*×*stbrush*	$-4\,649.81$	14.7	0
(6) *gender* + *caries* + *stbrush*	$-4\,645.49$	10.7	0

Table 3: Signal Tandmobiel Study. Estimates of regression parameters for models (1) and (2).

Covariate	Estimate (standard error; p-value)	exp(estimate)
	Model (1)	
gender	-0.173 (0.008; < 0.001)	0.841
	Model (2)	
gender	-0.173 (0.008; < 0.001)	0.841
caries	-0.029 (0.020; 0.148)	0.971

Table 4: WIHS Data. Akaike's information criterion, degrees of freedom, the optimal $\log(\lambda/n)$, estimates of the regression parameters (standard error; p-value) for the fitted models.

Model	AIC	df	$\log(\lambda)$	$dental$	$logvload$	$logcd4$
(1) $dental$	-262.43	3.2	2	-0.87 (0.36; 0.018)		
(2) $lvload$	-256.16	3.4	2		-0.76 (0.19; <0.001)	
(3) $lcd4$	-256.85	3.3	2			0.44 (0.11; <0.001)
(4) $dental + lvload$	-255.63	4.4	2	-0.62 (0.36; 0.080)	-0.70 (0.19; <0.001)	
(5) $dental + lcd4$	-253.19	8.9	-7	-0.78 (0.26; 0.003)		0.39 (0.07; <0.001)
(6) $lvload + lcd4$	-253.45	8.4	-6		-0.39 (0.14; 0.004)	0.38 (0.06; <0.001)
(7) $dental + lvload+$ $+lcd4$	-249.94	10.8	-8	-0.61 (0.21; 0.003)	-0.30 (0.10; 0.004)	0.39 (0.04; <0.001)