PENALIZED LIKELIHOOD IN COX REGRESSION

PIERRE J. M. VERWEIJ AND HANS C. VAN HOUWELINGEN

Department of Medical Statistics, Leiden University, P.O. Box 9604, 2300 RC Leiden, The Netherlands.

SUMMARY

In a Cox regression model, instability of the estimated regression coefficients can be reduced by maximizing a penalized partial log-likelihood, where a penalty function of the regression coefficients is substracted from the partial log-likelihood. In this paper, we choose the optimal weight of the penalty function by maximizing the predictive value of the model, as measured by the crossvalidated partial log-likelihood. Our methods are illustrated by a study of ovarian cancer survival and by a study of centre-effects in kidney graft survival.

1. INTRODUCTION

This paper deals with the modelling of categorical covariates in survival analysis, based on Cox's proportional hazards model.[1] The estimated regression coefficients of categorical covariates may be unstable, especially if some categories contain few observations. In linear models, parameter estimates can be improved by ridge regression, introduced by Hoerl and Kennard.[2] The ridge estimates are biased, but more stable than the conventional unbiased estimates. An equivalent method in likelihood based models is the estimation of coefficients by penalized likelihood, where a penalty function of the regression coefficients is subtracted from the log-likelihood. For examples in logistic regression see Le Cessie and Van Houwelingen[3] and Zwinderman *et al.*[4]

The main problem in estimation by penalized likelihood is how much weight to put on the penalty function. In this paper, the weight is determined by maximizing the predictive value of the model. In linear models, the predictive value can be measured by the predicted sum of squares PRESS.[5] In likelihood based models, this can be done by the crossvalidated log-likelihood CVL,[6] like PRESS computed from the leave-one-out regression coefficients. The CVL is related to Akaike's Information Criterion AIC,[7] but it has the advantage that it can also be used in Cox regression, where the components of the partial likelihood are dependent.

We apply these methods to categorical covariates in the Cox model. The penalized partial likelihood approach is outlined in Section 2. In Section 3, we define the CVL-measure of the predictive value. Applications of penalized likelihood in Cox regression are given in Sections 4 and 5. In Section 4, we optimize the coding of two ordinal covariates (categorical covariates with ordered categories) in ovarian cancer survival. Section 5 deals with the estimation of centre effects in kidney transplant survival.

2. PENALIZED PARTIAL LIKELIHOOD

Suppose we have n observations (t_i, d_i, X_i), where t_i is the possibly censored survival time, d_i is the censoring indicator and X_i is a row vector of covariates for individual i. In the Cox proportional

CCC 0277–6715/94/232427–10

hazards model,[1] the hazard function for individual i is given by

$$h_i(t) = h_0(t) \exp(X_i \beta)$$

where $h_0(t)$ is a baseline hazard function and $\exp(X_i\beta)$ is the relative risk or hazard ratio. Parameters are estimated by maximizing the partial log-likelihood

$$l(\beta) = \sum_{i=1}^{n} d_i \left(X_i \beta - \ln\left(\sum_{h \in R_i} \exp(X_h \beta) \right) \right)$$

where R_i is the set of all individuals at risk at time t_i. The maximum likelihood estimate of β is denoted by b.

We now define the penalized partial log-likelihood as

$$l^{\lambda}(\beta) = l(\beta) - \frac{1}{2}\lambda p(\beta)$$

where $p(\beta)$ is a penalty function and λ is a non-negative weight parameter, which, in this section, is considered as fixed. The factor $\frac{1}{2}$ is introduced for mathematical convenience. The value of β that maximizes the penalized partial likelihood depends on λ and is denoted by b^{λ}.

With only one categorical covariate with c categories, we write $X_i = (X_{i1}, \ldots, X_{ic})$, where $X_{ij} = I[X_i = j]$ is a dummy variable for category j with corresponding regression coefficient β_j. If the first category is considered as a baseline, we have $\beta_1 = 0$. If the categories of X are not ordered and if, in addition, we may assume that the effects of the various categories are not too different from the mean effect, a suitable penalty function is

$$p_0(\beta) = \sum_{j=1}^{c} (\beta_j - \bar{\beta})^2 .$$

This function penalizes β_j's that are furthest from the mean.

Covariates with ordered categories form a special class to which we will refer as ordinal covariates. If X is an ordinal covariate, the difference between the regression coefficients of two adjacent categories is supposed not too large. This leads to penalty functions that penalize first- or second-order differences of consecutive β_j's:

$$p_1(\beta) = \sum_{j=1}^{c-1} (\beta_{j+1} - \beta_j)^2$$

or

$$p_2(\beta) = \sum_{j=2}^{c-1} (\beta_{j+1} - 2\beta_j + \beta_{j-1})^2 .$$

It is informative to consider the null space of a penalty function, that is, the set of coefficients β for which $p(\beta) = 0$. For p_0 and p_1, the null space only contains $\beta_1 = \ldots = \beta_c = 0$, while for p_2, all linear sequences of β_j's that imply a linear covariate effect remain unpenalized as well. If there exists a prior belief about the behaviour of the regression coefficients, this can be reflected by an appropriate choice of the penalty function, especially its null space.

The extension to more covariates is straightforward. As an example we consider two ordinal covariates X_1 and X_2 with c_1 and c_2 categories and regression coefficients $\beta_j (j = 1, \ldots, c_1)$ and $\gamma_k (k = 1, \ldots, c_2)$, respectively. In an additive model the regression coefficient δ_{jk} of cell (j, k) can

be obtained as $\beta_j + \gamma_k$. If there is interaction between X_1 and X_2, all cells must be considered as separate categories. In both cases the δ_{jk}'s can be penalized by

$$p_3(\delta) = \sum_{j=1}^{c_1} \sum_{k=1}^{c_2} (\delta_{jk} - \bar{\delta}_{jk})^2$$

where $\bar{\delta}_{jk}$ is the mean of the δ_{jk}'s of the bordering categories.

Observe that coefficients of boundary cells are included in this function. They can be left out by summing from 2 to $c_1 - 1$ and $c_2 - 1$, respectively. The null space of p_3 is again only $\delta_{jk} = 0$ for all j and k.

All four penalty functions $p_0, \ldots, p_3$ can be written in the form $p(\beta) = \beta' A \beta$, with A a symmetric non-negative definite matrix. With $p(\beta)$ in this form the penalized log-likelihood can be written as

$$l^\lambda(\beta) = l(\beta) - \frac{1}{2}\lambda \beta' A \beta . \tag{1}$$

For a given λ, the first and second derivatives of the penalized log-likelihood with respect to β are given by

$$\frac{\partial l^\lambda}{\partial \beta}(\beta) = U^\lambda(\beta) = U(\beta) - \lambda A \beta = \frac{\partial l}{\partial \beta}(\beta) - \lambda A \beta$$

and

$$-\frac{\partial^2 l^\lambda}{\partial \beta^2}(\beta) = H^\lambda(\beta) = H(\beta) + \lambda A = -\frac{\partial^2 l}{\partial \beta^2}(\beta) + \lambda A,$$

respectively. A Newton–Raphson procedure can now be used to estimate the penalized regression coefficients b^λ.

A first-order Taylor expansion of U^λ around the unrestricted estimate b leads to the following approximation of b^λ:

$$b^\lambda = [H^\lambda(b)]^{-1} H(b) b . \tag{2}$$

In the Normal linear model, this is precisely the ridge estimate of Hoerl and Kennard,[2] if A is chosen as the identity matrix. Since b^λ is an intrinsically biased estimator, standard errors of b^λ are not very informative. Instead, we will report the square root of the diagonal elements of $[H^\lambda(b^\lambda)]^{-1}$ which give an impression of the stability of the penalized estimates. These quantities may be called pseudo-standard errors.

3. CROSSVALIDATED PARTIAL LIKELIHOOD

In this section, we determine the weight parameter, λ, by using the predictive value of the model as a criterion. Predictive value is conceptually different from explained variation; while the latter measures the fit to the data from which the model was derived, the predictive value is a measure for the fit to future data. As there are no future data, we mimic the prediction process by crossvalidation; every observation is left out once and predicted by using all other observations. In a Normal linear model with observations Y_i, the predictive value can be measured by the predicted sum of squares PRESS,[5] which is a function of the weight parameter λ:

$$\text{PRESS}(\lambda) = \sum_{i=1}^{n} (Y_i - X_i b^\lambda_{(-i)})^2 .$$

Here $b^{\lambda}_{(-i)}$ denotes the 'leave-one-out' regression coefficient, that is estimated when individual i is left out of the observations.

In a likelihood based (Cox) model, a generalization of PRESS is the crossvalidated (partial) log-likelihood[6]

$$\mathrm{CVL}(\lambda) = \sum_{i=1}^{n} l_i(b^{\lambda}_{(-i)})$$

where $l_i(\beta)$ is the contribution of individual i to the log-likelihood, defined as $l(\beta) - l_{(-i)}(\beta)$, with $l_{(-i)}(\beta)$ denoting the unrestricted log-likelihood without individual i. The value of λ that maximizes $\mathrm{CVL}(\lambda)$ is the optimal weight parameter to use in the penalized (partial) log-likelihood.

In the Cox model, the following expression for $l_i(\beta)$ can be derived:[6]

$$l_i(\beta) = \ln \prod_{t_k < t_i} [(1 - p_{ik})^{d_k} p_{ik}^{d_{ki}}]$$

with

$$p_{ik} = \frac{\exp(X_i \beta)}{\sum_{k \in R_k} \exp(X_k \beta)}$$

the conditional probability that individual i dies at time t_k, given the individual is alive just before t_k. Observe that for $l_i(\beta)$ the product only includes the observed failure times before t_i. Hence, $l_i(\beta)$ can be interpreted as the (log) probability that individual i survives at all occasions before t_i and dies at t_i.

In practice we will use the following approximation to the CVL:[6]

$$\mathrm{CVL}(\lambda) = \sum_{i=1}^{n} l_i(b^{\lambda}) - c(\lambda) \tag{3}$$

with

$$c(\lambda) = tr\left([H^{\lambda}(b^{\lambda})]^{-1} \sum_{i=1}^{n} U_i(b^{\lambda})\, U_i(b^{\lambda})^T \right) \tag{4}$$

a term that corrects the sum of the individual contributions. Here $U_i(\beta)$ denotes the vector of derivatives of $l_i(\beta)$ with respect to β.

If the components of the log-likelihood are independent, as in linear and logistic models, the first term in the CVL equals $l(b^{\lambda})$, the log-likelihood of the model evaluated at b^{λ}, while $c(\lambda)$ in expectation equals the effective dimension

$$e(\lambda) = \mathrm{tr}[H^{\lambda}(b^{\lambda})]^{-1} H(b^{\lambda}). \tag{5}$$

Hence, with independent components, CVL (λ) is approximately equal to a penalized version of Akaike's Information Criterion[7]

$$\mathrm{AIC}(\lambda) = l(b^{\lambda}) - e(\lambda). \tag{6}$$

Observe that AIC(0) is the original definition of AIC, that is, the log-likelihood minus the dimension of the model.

In the Cox model, the components $l_i(\beta)$ are dependent, which implies that the sum of the $l_i(\beta)$'s is not equal to $l(\beta)$ and AIC (λ) is not a numerical approximation to CVL (λ). However, the differences ΔAIC and ΔCVL from the null model (without covariates) are approximately equal. Hence, AIC(λ) can be useful as an alternative criterion for the determination of the weight parameter in penalized partial likelihood.

Table I. Diameter of residual tumour in ovarian cancer survival. Categories, number of patients (n) and regression coefficients estimated without penalty (b with standard error) and with penalty functions p_1 (b_1 with pseudo-standard error, PSE) and p_2 (b_2 with PSE). Correction c, from (4), and CVL, from (3), are given at the bottom of the table

Category	Residual diameter	n	b	(SE)	b_1	(PSE)	b_2	(PSE)
1	microsc	29	0		0		0	
2	< 1cm	67	0·424	(0·319)	0·233	(0·189)	0·306	(0·093)
3	1–2 cm	49	0·934	(0·322)	0·606	(0·220)	0·600	(0·146)
4	2–5 cm	68	1·023	(0·310)	0·781	(0·218)	0·861	(0·174)
5	> 5cm	145	1·234	(0·293)	0·965	(0·207)	1·105	(0·198)
	c			3·91		2·34		1·19
	CVL			− 1664·57		− 1663·65		− 1662·68

Table II. Karnofsky performance status in ovarian cancer survival. Categories, number of patients (n) and regression coefficients estimated without penalty (b with standard error) and with penalty functions p_1 (b_1 with pseudo-standard error, PSE) and p_2 (b_2 with PSE). Correction c, from (4), and CVL, from (3), are given at the bottom of the table

Category	Performance status	n	b	(SE)	b_1	(PSE)	b_2	(PSE)
1	100	137	0		0		0	
2	90	108	0·071	(0·155)	0·093	(0·123)	0·198	(0·090)
3	80	47	0·314	(0·196)	0·350	(0·152)	0·466	(0·129)
4	70	46	0·811	(0·188)	0·715	(0·160)	0·802	(0·153)
5	⩽ 60	20	1·169	(0·253)	0·921	(0·209)	1·156	(0·210)
	c			4·57		2·49		1·71
	CVL			− 1668·49		− 1667·12		− 1666·28

4. OPTIMAL CODING OF ORDINAL COVARIATES IN OVARIAN CANCER SURVIVAL

We consider the survival times for 358 patients with an advanced ovarian cancer, who were treated with chemotherapy. The data were derived from two trials comparing two different types of chemotherapy.[8] The follow-up was at least 4 years and the 4 year survival probability was about 0·30. We concentrate on the coding of two ordinal covariates: the diameter of the residual tumour (after surgery) and the Karnofsky performance status (measured at start of treatment), which is a measure of the ability of a patient to lead her daily life. Both covariates have five categories, which are defined in Tables I and II. Throughout we assume a proportional hazards model. In a notation slightly different from Section 2, we write the hazard function as $h(t) = h_0(t)$ RR, where RR is the relative risk or hazard ratio depending on one or both covariates.

If there is only one ordinal covariate X with c categories, the relative risk can be modelled as

$$\ln(\mathrm{RR}) = \sum_{j=1}^{c} \beta_j I[X = j].$$

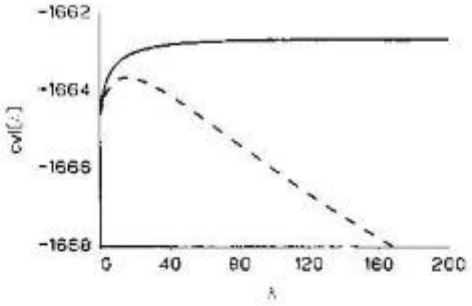

Figure 1. Diameter of residual tumour in ovarian cancer survival. Crossvalidated partial log-likelihood CVL(λ) plotted against λ for models with penalty functions p_1 (dashed) and p_2 (solid)

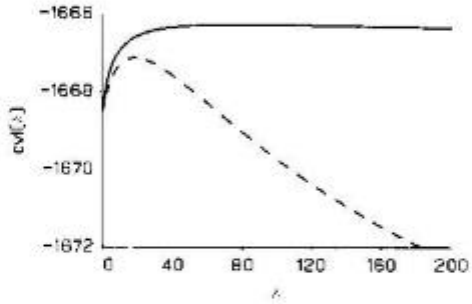

Figure 2. Karnofsky performance status in ovarian cancer survival. Crossvalidated partial log-likelihood CVL(λ) plotted against λ for models with penalty functions p_1 (dashed) and p_2 (solid)

For the residual diameter and karnofsky index, we have $c = 5$ and the estimated regression coefficients are given in Tables I and II, respectively. The stability of the coefficients and the predictive value of the model can be improved by estimating the regression coefficients by using the penalized partial log-likelihood (1). As penalty functions we consider p_1 and p_2 of Section 2. The value of the weight parameter is based on the maximization of the approximate CVL (3).

First we consider a model with the residual diameter as the only covariate. For $\lambda = 0$, the correction (4), is about 4, which is the dimension of the unrestricted model. As $\lambda \to \infty$, $c(\lambda) \to 0$ for p_1 and $c(\lambda) \to 1$ for p_2, these values coinciding with the dimension of the null spaces of the penalty functions. The crossvalidated log-likelihood CVL(λ) is plotted in Figure 1. It is clear that penalty function p_2 leads to a higher predictive value than p_1. For p_1, the optimal weight parameter is 15·4, while for p_2, the optimal λ seems to be ∞. In the latter case, this would lead to a linear model for residual diameter. However, for λ about 176, the CVL is slightly higher than the CVL of the linear model ($-1662{\cdot}68$ versus $-1662{\cdot}72$). The penalized estimates of the regression coefficients are given in Table I.

For the Karnofsky performance status, the crossvalidated log-likelihood is plotted in Figure 2. Again, the use of penalty function p_2 leads to the highest predictive value. Furthermore, for λ near 75, the CVL is a bit higher than the CVL of the linear model ($-1666{\cdot}28$ versus $-1666{\cdot}62$). The estimated regression coefficients are given in Table II.

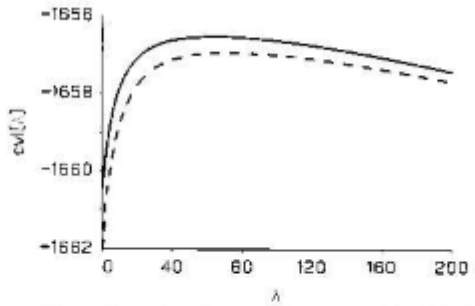

Figure 3. Residual diameter and Karnofsky index. Crossvalidated partial log-likelihood CVL(λ) plotted against λ for an additive model (solid) and for the model including interaction (dashed), both with penalty function p_3

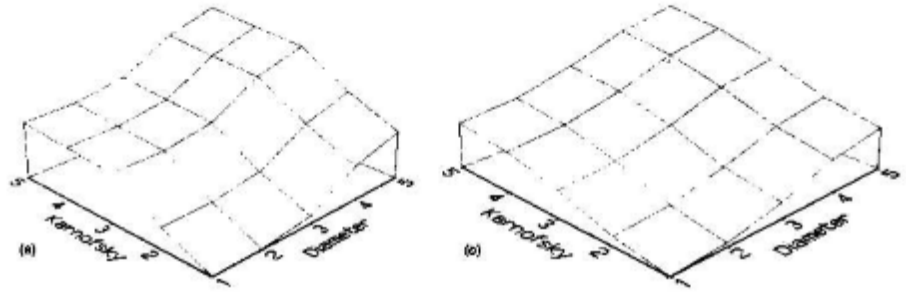

Figure 4. Residual diameter and Karnofsky index. The ln (relative risk) surfaces arising from an additive model: (*a*) unrestricted (*b*) penalized

With two ordinal covariates X_1 and X_2, with c_1 and c_2 categories and regression coefficients $\beta_j (j = 1 \ldots c_1)$ and $\gamma_k (k = 1 \ldots c_2)$, respectively, for which the effects are additive, we model the relative risk as follows:

$$\ln(\mathrm{RR}) = \sum_{j=1}^{c_1} \beta_j I[X_1 = j] + \sum_{k=1}^{c_2} \gamma_k I[X_2 = k] .$$

Since the right hand side is of the form $f_1(X_1) + f_2(X_2)$, with f_1 and f_2 possibly non-linear functions, this model is a generalized additive model in the spirit of Hastie and Tibshirani.[9] With X_1 representing residual diameter and X_2 the Karnofsky index, we have $c_1 = c_2 = 5$ and the ln(RR) arising from the unrestricted estimates, that varies from 0 to 1·905, is plotted in Figure 4(*a*).

Again, we use penalized likelihood estimation to improve the predictive value, the penalty function being p_3 of Section 2, where $\delta_{jk} = \beta_j + \gamma_k$. The CVL is plotted in Figure 3. The optimal value of λ is about 64. The corresponding ln(RR) varies from 0 to 1·485 and is plotted in Figure 4(*b*). This surface is smoother than the one in Figure 4(*a*) and suggests a model in which both covariates are linear. The predictive value of the linear model, however, is a bit lower (CVL −1657·04 versus −1656·51).

It is also possible to consider a saturated model by adding an interaction term between residual diameter and Karnofsky index to the model:

$$\ln(\mathrm{RR}) = \sum_{j=1}^{c_1} \sum_{k=1}^{c_2} \delta_{jk} I[X_1 = j, X_2 = k] .$$

In this case, every combination of the two covariates is treated as a separate category. The unrestricted estimates of the regression coefficients are highly unstable, because 11 of the 25 categories contain less than 10 observations, including two empty categories. Penalized estimation of the regression coefficients is now the only way to obtain realistic values. The penalty function is again p_3 of Section 2.

The predictive value, as measured by CVL(λ), is plotted in Figure 3. It can be seen that for every value of λ, CVL(λ) of the interaction model is lower than CVL(λ) of the additive model. Hence, adding the interaction term does not improve the predictive value, neither for the unrestricted model, nor for any penalized model. The maximal CVL of the interaction model ($-1656{\cdot}95$) is obtained when λ is about 72. The corresponding ln(RR) values, ranging from 0 to $1{\cdot}454$, are approximately the same as those arising from the additive model.

Another approach to the two-covariate situation is the use of separate penalty functions $p(\beta)$ and $q(\gamma)$, with possibly different weight parameters λ and μ, leading to the penalized log-likelihood:

$$l^{\lambda,\mu}(\beta, \gamma) = l(\beta, \gamma) - \tfrac{1}{2}\lambda p(\beta) - \tfrac{1}{2}\mu q(\gamma).$$

In view of the previous results, function p_2 from Section 2 is appropriate for both p and q. The results of this approach (not shown) do not differ much from the reported results, but the computations are more time consuming, because the CVL is in this case a function of two arguments.

5. CENTRE EFFECTS IN KIDNEY GRAFT SURVIVAL

An example of a categorical covariate without ordered categories is the centre effect in a multicentre study. Here we consider data from 4754 (first) kidney transplants, taken from the database of the Eurotransplant foundation and analysed by Thorogood *et al.*[10] The transplantations were performed in 52 centres in the years 1984–1987. We left out five observations from three centres with only one or two transplantations. For the remaining 49 centres, the number of transplantations varied from 7 to 391. The centre effects were coded as a categorical covariate with 49 categories. We did not choose a baseline centre, but constrained the estimated coefficients to sum to zero. Apart from the centre effects, nine other covariates were included in the model: donor and recipient sex; donor age; recipient age; cold ischaemic period; the number of HLA-DR mismatches; the number of HLA-B mismatches; highest panel reactive antibody; recipient blood group and recipient diabetic. The unrestricted estimates of the centre effects are plotted against the rank of their standard error in Figure 5. For small centres, the standard errors are large and the estimates are rather unstable, as can be seen in the right hand part of the figure.

In the penalized estimation procedure, an appropriate penalty function for the centre effects is p_0 from Section 2. This function can be written as $\beta^T(I - P)\beta$, where I is the (49×49) identity matrix and P is a matrix with $1/49$ in every entry. The regression coefficients of the other nine covariates remain unpenalized. For the determination of the weight parameter, we use AIC(λ) (6) instead of CVL(λ), because in this large dataset the former is computationally more attractive. AIC(λ) is maximized by $\lambda = 3{\cdot}0$ and the corresponding penalized estimates of the centre effects are plotted in Figure 5. For the small centres in the right of the picture, a shrinkage phenomenon can be observed; the positive estimates have decreased and the negative estimates have increased. The parameter estimates of the nine other covariates (not shown), which were not penalized, changed only very slightly, as expected.

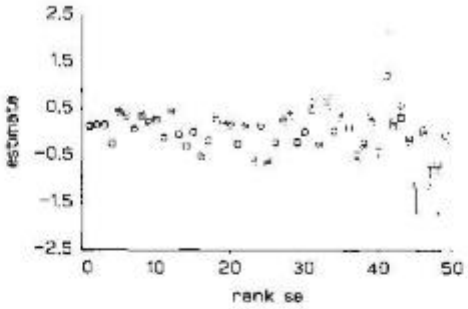

Figure 5. Centre effects in kidney transplantation. Unrestricted (circles) and penalized estimates (squares) plotted against the rank of the standard error of the unrestricted effects

A different way to obtain an estimate of λ is from the empirical Bayes perspective. The penalized partial log-likelihood

$$l^{\lambda}(\beta) = l(\beta) - \frac{1}{2}\lambda\beta^{T}(I - P)\beta$$

can be seen as the unnormalized posterior log-likelihood of β with respect to the Normal prior

$$\beta \sim \mathrm{N}\left(0, \frac{1}{\lambda}(I - P)^{-}\right).$$

To convert it into a true posterior, a λ-dependent normalizing term has to be added. Observe that the covariance matrix is singular because of the constraint $\Sigma\beta_j = 0$. Hence, a generalized inverse is used instead of an ordinary inverse. Our b^{λ} is the posterior mode of β. If the conditional distribution of b given β is supposed to be $N(\beta, \Sigma_b)$ with $\Sigma_b = H(b)^{-}$, the posterior mode equals the posterior mean and is given by

$$\frac{1}{\lambda}(I - P)^{-}\left[H(b)^{-} + \frac{1}{\lambda}(I - P)^{-}\right]^{-} b = [H(b) + \lambda(I - P)]^{-} H(b)b$$

which is the same expression as the approximation (2) of b^{λ}, given in Section 2. Now λ can be estimated from the marginal distribution of the data. Under the Normal approximation, we have

$$Eb'b = \mathrm{tr}[H(b)^{-}] + \frac{1}{\lambda}\mathrm{tr}(I - P)^{-}$$

leading to the moment estimation

$$\hat{\lambda} = \frac{\mathrm{tr}(I - P)^{-}}{b'b - \mathrm{tr}[H(b)^{-}]}$$

in the spirit of DerSimonian and Laird.[11] This gives a value of 3·9 for the optimal λ. A more sophisticated estimate can be obtained from the ML estimate of λ in the marginal model, which can be obtained by EM methodology,[12] but that is beyond the scope of this paper. Finally, we remark that the posterior covariance matrix of β is approximately equal to $H^{\lambda}(b)^{-}$, which motivates the pseudo standard errors of Section 2 as Bayesian posterior standard deviations.

6. DISCUSSION

Penalized partial likelihood is a useful method for obtaining more stable estimates of regression coefficients for Cox's proportional hazards model. For ordinal covariates the penalized estimates lead to smoother surfaces of log (relative risk), while for covariates without ordered categories, like centre effects, the penalized estimates are shrunk towards the mean. Although we restricted ourselves to categorical covariates, the penalized partial likelihood method can also be applied to continuous covariates which are highly collinear. If the predictive value of the model is used as a criterion to determine the weight parameter, the penalized estimates have the extra advantage of being the best predictors.

Penalized (partial) likelihood estimation is related to estimation in frailty models[13,14] in which the centre effects are considered as random. This is also the case in empirical Bayes estimation, which allows the weight parameter to be estimated from the marginal distribution of the data. Using a Normal approximation to $l(\beta)$, a moment type estimator for λ can be obtained that is very simple to compute. However, it can be conjectured that this estimate is not robust against deviation from the prior model, while the crossvalidated approach does not need such an assumption and can be expected to be robust against unrealistic penalties.

Penalized likelihood methods are not incorporated in standard software for survival analysis, but a program written in GAUSS is available from the first author.

REFERENCES

1. Cox, D. R. 'Regression models and life tables' (with discussion), *Journal of the Royal Statistical Society, Series B*, **34**, 187–220 (1972).
2. Hoerl, A. E. and Kennard, R. W. 'Ridge regression: Biased estimates for nonorthogonal problems', *Technometrics*, **12**, 55–67 (1970).
3. Le Cessie, S. and Van Houwelingen, J. C. 'Ridge estimators in logistic regression', *Applied Statistics*, **41**, 191–201 (1992).
4. Zwinderman, A. H., Van Holten-Verzantvoort, A. T. M., Aaronson, N. K. and Van Houwelingen, J. C. 'The analysis of repeated quality of life measurements with an autoregressive latent trait model', *Journal of Quality of Life Research*, **1**, 219–224 (1992).
5. Allen, D. M. 'The relation between variable selection and data augmentation and a method for prediction', *Technometrics*, **16**, 125–127 (1974).
6. Verweij, P. J. M. and Van Houwelingen, J. C. 'Crossvalidation in survival analysis', *Statistics in Medicine*, **12**, 2305–2314 (1993).
7. Akaike, H. 'Information theory and an extension of the entropy maximization principle', *in* Petrov, B.N. and Csak, F. (eds), *2nd International Symposium on Information Theory*, Akdemia, Kiado, 1973, pp. 267–281.
8. Van Houwelingen, J. C., Ten Bokkel Huinink, W. W., Van der Burg, M. E. L., Van Oosterom, A. T. and Neijt, J. P. 'Predictability of the survival of patients with advanced ovarian cancer', *Journal of Clinical Oncology*, **7**, 769–773 (1989).
9. Hastie, T. J. and Tibshirani, R. J. 'Generalized additive models' (with discussion), *Statistical Science*, **1**, 297–318 (1986).
10. Thorogood, J., Van Houwelingen, J. C., Persijn, G. G., Zantvoort, F. A., Schreuder, G. M. Th. and Van Rood, J. J. 'Prognostic indices for first and second transplant can predict kidney graft survival', *Transplantation*, **52**, 831–836 (1991).
11. DerSimonian, R. and Laird, N. M. 'Meta-analysis in clinical trials', *Controlled Clinical Trials*, **7**, 177–188 (1986).
12. Klein, J. P. 'Semiparametric estimation of random effects using the Cox model based on the EM algorithm', *Biometrics*, **48**, 795–806 (1992).
13. McGilchrist, C. A. and Aisbett, C. W. 'Regression with frailty in survival analysis', *Biometrics*, **47**, 461–466 (1991).
14. McGilchrist, C. A. 'REML estimation for survival models with frailty', *Biometrics*, **49**, 221–225 (1993).