STATISTICS IN MEDICINE, VOL. 16, 385–395 (1997)

THE LASSO METHOD FOR VARIABLE SELECTION IN THE COX MODEL

ROBERT TIBSHIRANI

Department of Preventive Medicine and Biostatistics and Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 1A8

SUMMARY

I propose a new method for variable selection and shrinkage in Cox's proportional hazards model. My proposal minimizes the log partial likelihood subject to the sum of the absolute values of the parameters being bounded by a constant. Because of the nature of this constraint, it shrinks coefficients and produces some coefficients that are exactly zero. As a result it reduces the estimation variance while providing an interpretable final model. The method is a variation of the 'lasso' proposal of Tibshirani, designed for the linear regression context. Simulations indicate that the lasso can be more accurate than stepwise selection in this setting.

1. INTRODUCTION

Consider the usual survival data setup. The date available are of the form $(y_1, \mathbf{x}^1, \delta_1), \ldots, (y_N, \mathbf{x}^N, \delta_N)$, the survival time y_i being complete if $\delta_i = 1$ and right censored if $\delta_i = 0$, with $\mathbf{x}^i$ denoting the usual vector of predictors $(x_1, x_2, \ldots, x_p)$ for the ith individual. Denote the distinct failure times by $t_1 < \cdots < t_k$, there being d_i failures at time t_i.

The proportional-hazards model for survival data, also known as the Cox model,[1] assumes that

$$\lambda(t|\mathbf{x}) = \lambda_0(t) \exp\left(\sum_j x_j \beta_j\right) \tag{1}$$

where $\lambda(t|\mathbf{x})$ is the hazard at time t given predictor values $\mathbf{x} = (x_1, \ldots, x_p)$, and $\lambda_0(t)$ is an arbitrary baseline hazard function.

One usually estimates the parameter $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^{\mathrm{T}}$ in the proportional-hazards model (1) without specification of $\lambda_0(t)$ through maximization of the the partial likelihood

$$L(\boldsymbol{\beta}) = \prod_{r \in D} \frac{\exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}^{j_r})}{\{\sum_{j \in R_r} \exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}^j)\}} . \tag{2}$$

In equation (2) D is the set of indices of the failures, R_r is the set of indices of the individuals at risk at time $t_r - 0$, and j_r is the index of the failure at time t_r. Assume for simplicity that there are no tied failure times; suitable modifications of the partial likelihood exist for the case of ties. Assume also that the censoring is non-informative, so that the construction of the partial likelihood is justified.

CCC 0277–6715/97/040385–11

Received March 1995
Revised December 1995

Denote the log partial likelihood by $\ell(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta})$, and assume that the x_{ij} are standardized so that $\sum_i x_{ij}/N = 0$, $\sum_i x_{ij}^2/N = 1$.

In this paper I propose to estimate $\boldsymbol{\beta}$ via the criterion

$$\hat{\boldsymbol{\beta}} = \text{argmin}\, \ell(\boldsymbol{\beta}), \quad \text{subject to } \sum |\beta_j| \leqslant s \tag{3}$$

where $s > 0$ is a user-specified parameter. In the linear regression setting, Tibshirani[2] proposed minimization of the residual sum of squares, subject to a constraint of the form $\sum |\beta_j| \leqslant s$ and called the resulting procedure the 'lasso' for 'Least Absolute Shrinkage and Selection Operator'. Here I use the term 'lasso' for the present proposal as well.

Suppose $\hat{\beta}_j^o$ are the maximizers of the partial likelihood (2). Then if $s \geqslant \sum |\hat{\beta}_j^o|$, the solutions to (3) are the usual partial likelihood estimates. If $s < \sum |\hat{\beta}_j^o|$, however, then the solutions to (3) are shrunken towards zero. An attractive feature of the particular constraint $\sum |\beta_j| \leqslant s$ is that quite often some of the solution coefficients are exactly zero. This makes for a more interpretable final model. On the other hand, the smooth form of the constraint should provide a more stable final model than that given by stepwise or best subset selection. In the regression setting, Tibshirani[2] confirmed this in simulation studies. In contrast, the ridge regression approach (used mainly in the linear model setting) shrinks coefficients but does not give coefficients that are exactly zero. Note that like model selection, the lasso is a tool for achieving parsimony; in actuality an exact zero coefficient is unlikely to occur.

The next section gives an algorithm for obtaining the lasso estimates. Section 3 contains two real data examples. Automatic estimation of the constraint parameter s appears in Section 4, and in Section 5 I report a simulation study that compares the lasso to stepwise selection. Section 6 discusses estimation of standard errors, while Section 7 contains some discussion, including a brief summary of other approaches to model selection.

2. COMPUTATION OF THE ESTIMATES

Tibshirani[2] gave two algorithms for the lasso procedure in the least squares regression setting, based on quadratic programming techniques. The algorithms are iterative and involve repeated solution of least squares problems. Typically one needs between p and $2p$ iterations, where p is the number of regressor variables.

The strategy for solving (3) is to express the usual Newton–Raphson update as an iterative reweighted least squares (IRLS) step, and then replace the weighted least squares step by a constrained weighted least squares procedure. If $\mathbf{X}$ denotes the design matrix of regressor variables and $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, define $\mathbf{u} = \partial\ell/\partial\boldsymbol{\eta}$, $\mathbf{A} = -\partial^2\ell/\partial\boldsymbol{\eta}\boldsymbol{\eta}^{\mathrm{T}}$ and $\mathbf{z} = \boldsymbol{\eta} + \mathbf{A}^{-1}\mathbf{u}$ (detailed expressions for $\mathbf{u}, \mathbf{A}$ and $\mathbf{z}$ appear in Hastie and Tibshirani,[3] Chapter 8, pp. 213–214). Then a one-term Taylor series expansion for $\ell(\beta)$ has the form

$$(\mathbf{z} - \boldsymbol{\eta})^{\mathrm{T}}\, \mathbf{A}\, (\mathbf{z} - \boldsymbol{\eta})\,. \tag{4}$$

Hence to solve the original problem (3), we use the following procedure:

1. Fix s and initialize $\hat{\boldsymbol{\beta}} = 0$.
2. Compute $\boldsymbol{\eta}$, $\mathbf{u}$, $\mathbf{A}$ and $\mathbf{z}$ based on the current value of $\hat{\boldsymbol{\beta}}$.
3. Minimize $(\mathbf{z} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{A}(\mathbf{z} - \mathbf{X}\boldsymbol{\beta})$ subject to $\sum |\beta_i| \leqslant s$.
4. Repeat steps 2 and 3 until $\hat{\boldsymbol{\beta}}$ does not change.

The minimization in step 3 is done through a quadratic programming procedure, as described in Tibshirani.[2] Note that if one used instead an unconstrained minimization in step 3, this procedure

would be equivalent to the usual Newton–Raphson algorithm for maximizing the partial likelihood (Hastie and Tibshirani,[3] Chapter 8, p. 212).

One difficulty with the above procedure is that **A** is a full matrix and hence it requires computation of $O(N^2)$ elements. One can avoid this, however by, replacing **A** with a diagonal matrix **D** that has the same diagonal elements as **A**. As argued in Hastie and Tibshirani[3] (Chapter 8, pp. 212–213), the diagonal elements of **A** are larger than the off-diagonal elements and hence the modified algorithm should behave similarly to the original one.

There is no intercept in the Cox model and hence the minimization in step 3 does not require an intercept. I have found however that inclusion of an intercept dramatically improves the convergence of the procedure. This is purely a computational issue; it makes no difference in the final model, as the intercept is absorbed into the baseline hazard.

If the log partial likelihood is bounded in $\boldsymbol{\beta}$ for the given data set, then for fixed s a solution to (3) exists since the region $\sum|\beta_j| \leqslant s$ is compact. However, the solution may not be unique. For example, if two regressors variables X_1 and X_2 are identically equal, then if $\beta > 0$, for any γ in $[0, \beta]$ the linear combination $X_1\gamma + (\beta - \gamma)X_2$ has exactly the same value for the ℓ and the constraint $|\gamma| + |\beta - \gamma|$. Note that this is due to the linearity of the constraint; in ridge style penalization involving the squared coefficients, this does not occur.

3. EXAMPLES

3.1. Lung cancer data

The data in this example come from the Veteran's Administration lung cancer trial, listed in Kalbfleisch and Prentice,[4] pp. 223–224. The time variable is survival in days, and the regressors are:

1. Treatment 1 = standard, 2 = test.
2. Cell type 1 = squamous, 2 = small cell, 3 = adeno, 4 = large.
3. Karnofsky score.
4. Months from diagnosis.
5. Age in years.
6. Prior therapy 0 = no, 10 = yes.

For simplicity, and because the categories exhibit increasing risk, I have left cell type as a numerical variable. A standard proportional hazards analysis shows that the Karnofsky score is extremely important, while cell type is also strongly significant. The other regressors show moderate effects.

Figure 1 shows the estimated coefficients from the lasso fit as a function of the standardized constraint parameter $u = s/\sum|\hat{\beta}_j^o|$ (where β_j^o are the unconstrained partial likelihood estimates). Karnofsky score is clearly the dominant effect with treatment and cell type also showing moderate influence. The vertical broken line is drawn at 0·45, which is the value of u chosen by generalized cross-validation (GCV, see Section 4). The model selected by generalized cross-validation has a non-zero coefficient only for Karnofsky score, with a coefficient of $-0{\cdot}47$, corresponding to a relative risk of 0·63. It standard error is 0·085, computed by the technique discussed in Section 6. Backward stepwise selection in the standard Cox model yields the same single variable model, but with a coefficient of $-0{\cdot}67$ (0·10) or a relative risk 0·51. The stepwise method refers to backward–forward stepwise selection as implemented in Scott Emerson's S language function 'coxrgrss' with the default P-values to enter and remove of 0·05 and 0·10, respectively. I also applied Schwarz's criterions (also known as BIC) to these data; this has the

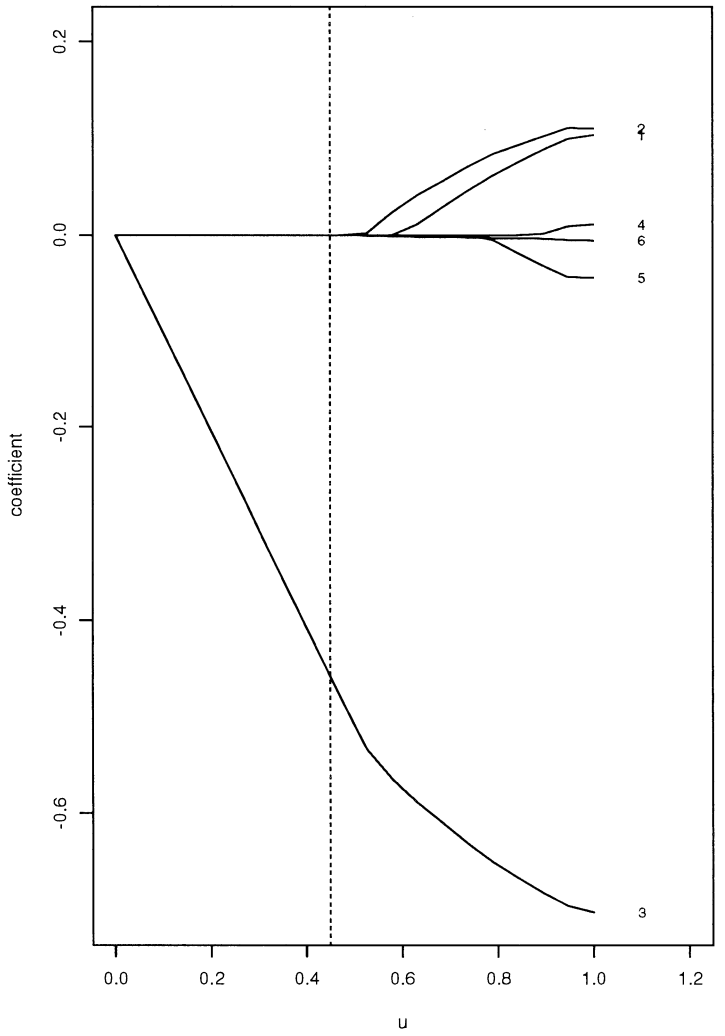

Figure 1. Coefficient estimates for lung cancer example, as a function of the standardized constraint parameter $u = s/\sum|\hat{\beta}_j^o|$

form minus log partial likelihood plus $k\log n$ where k is the number of regressors in the model considered and n is the sample size. Searching over all subsets, the model that minimizes Schwarz's criterion again contained only the Karnofsky score.

3.2. Liver data

The data in this example and the following (edited) description were provided by Harrington and Fleming.

> 'Primary biliary cirrhosis (PBC) of the liver is a rare but fatal chronic liver disease of unknown cause, with a prevalence of about 50-cases-per-million population. The primary pathologic event appears to be the destruction of interlobular bile ducts, which may be mediated by immunologic mechanisms.

> The following briefly describes data collected for the Mayo Clinic trial in PBC of the liver conducted between January 1974 and May 1984 comparing the drug D-penicillamine (DPCA) with a placebo. The first 312 cases participated in the randomized trial of D-penicillamine versus placebo, and contain largely complete data. An additional 112 cases did not participate in the clinical trial, but consented to have basic measurements recorded and to be followed for survival. Six of those cases were lost to follow-up shortly after diagnosis, so there are data here on an additional 106 cases as well as the 312 randomized participants.'

I discarded observations with missing values, leaving 276 observations. The variables in the data set are:

N Case number.
Y The number of days between registration and the earlier of death or study analysis time in 1986.
δ 1 if Y is time to death, 0 if time to censoring.
X_1 Treatment code, 1 = D-pencillamine, 2 = placebo.
X_2 Age in years. For the first 312 cases, age was calculated by dividing the number of days between birth and study registration by 365.
X_3 Sex, 0 = male, 1 = female.
X_4 Presence of ascites, 0 = no, 1 = yes.
X_5 Presence of hepatomegaly, 0 = no, 1 = yes.
X_6 Presence of spiders, 0 = no, 1 = yes.
X_7 Presence of oedema, 0 = no, 0·5 = yes but responded to diuretic treatment, 1 = yes, did not respond to treatment.
X_8 Serum bilirubin, in mg/dl.
X_9 Serum cholesterol, in mg/dl.
X_{10} Albumin, in g/dl.
X_{11} Urine copper, in μg/day.
X_{12} Alkaline phosphatase, in U/litre.
X_{13} SGOT, in U/ml.
X_{14} Triglycerides, in mg/dl.
X_{15} Platelet count; coded value is number of platelets per cubic ml of blood divided by 1000.
X_{16} Prothrombine time, in seconds.
X_{17} Histologic state of disease, graded 1, 2, 3 or 4.

Some results appear in Table I. The stepwise method refers to backward–forward stepwise selection as implemented in Scott Emerson's S language function 'coxrgrss' with the default P-values to enter and remove of 0·05 and 0·10 respectively. Coxrgrss is available from the Statlib archive (ftp site lib.stat.cmu.edu). This gave a model with eight variables, all having large Z-scores. The GCV procedure gave $\hat{u} = 0{\cdot}56$ for the standardized lasso parameter and the resulting model from the lasso looks similar to the stepwise model, with most of the effects shrunken towards zero. While the stepwise procedure often inflates the Z scores of chosen variables relative to the full model fit, the lasso seems to shrink them towards zero. I computed standard errors for the lasso estimates using the method given in Section 6.

4. ESTIMATION OF THE CONSTRAINT PARAMETER s

In some situations it is desirable to have an automatic method for choosing s based on the data. Such a procedure is analogous to an automatic subset selection procedure such as forward, backward or all subsets regression.

Table I. Results for liver data example

Variables	Full			Stepwise			Lasso		
	Coefficient	SE	Z-score	Coefficient	SE	Z-score	Coefficient	SE	Z-score
1	− 0·06	0·11	− 0·58	–	–	–	0·00	0·00	0·00
2	0·30	0·12	2·49	0·33	0·11	3·08	0·17	0·09	1·89
3	− 0·12	0·10	− 1·17	–	–	–	− 0·01	0·03	− 0·31
4	0·02	0·10	0·23	–	–	–	0·04	0·07	0·63
5	0·01	0·13	0·10	–	–	–	0·00	0·00	0·00
6	0·05	0·11	0·42	–	–	–	0·02	0·05	0·40
7	0·27	0·11	2·56	0·22	0·09	2·37	0·18	0·11	1·71
8	0·37	0·12	3·14	0·39	0·09	4·39	0·35	0·12	2·97
9	0·12	0·10	1·11	–	–	–	0·00	0·01	0·28
10	− 0·30	0·12	− 2·40	− 0·29	0·11	− 2·63	− 0·22	0·10	− 2·27
11	0·22	0·10	2·13	0·25	0·09	2·90	0·21	0·11	1·98
12	0·00	0·08	0·03	–	–	–	0·00	0·00	0·00
13	0·23	0·11	2·08	0·25	0·10	2·42	0·09	0·08	1·04
14	− 0·06	0·09	− 0·75	–	–	–	0·00	0·00	0·00
15	0·08	0·11	0·76	–	–	–	0·00	0·00	0·00
16	0·23	0·11	2·19	0·23	0·10	2·25	0·09	0·09	0·97
17	0·39	0·15	2·59	0·37	0·12	2·97	0·21	0·09	2·28

My proposal is to minimize an approximate generalized cross-validation (GCV) statistic (Wahba[6]). To construct this statistic, we need a linear approximation to the lasso estimate. We write the constraint $\sum|\beta_j| \leqslant s$ as $\sum \beta_j^2/|\beta_j| \leqslant s$. This latter constraint is equivalent to adding a Lagrangian penalty $\lambda \sum \beta_j^2/|\beta_j|$ to the log partial likelihood, with $\lambda \geqslant 0$ depending on s. Intuitively, these are equivalent since they both lead to a balance between fit, as measured by the log partial likelihood, and the value of $\lambda \sum \beta_j^2/|\beta_j|$. Using standard matrix manipulations, we may write the constrained solution $\tilde{\boldsymbol{\beta}}$ in step 3 in the form

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}} \mathbf{D} \mathbf{X} + \lambda \mathbf{W})^{-1} \mathbf{X}^{\mathrm{T}} \mathbf{D} \mathbf{z} \tag{5}$$

$\mathbf{W} = \mathrm{diag}(\mathbf{W}_j)$, $\mathbf{W}_i = 1/|\tilde{\beta}_j|$ if $|\tilde{\beta}_j| > 0$ and 0 otherwise. (This expression does not give a numerically effective way of computing the lasso estimate, but it is useful for assessing the complexity of the fit.) Therefore we may approximate the number of effective parameters in the constrained fit $\tilde{\boldsymbol{\beta}}$ by

$$p(s) = \mathrm{tr}\left[\mathbf{X}(\mathbf{X}^{\mathrm{T}} \mathbf{D} \mathbf{X} + \lambda \mathbf{W}^{-})^{-1} \mathbf{X}^{\mathrm{T}} \mathbf{D}\right].$$

Letting ℓ_s be the log-partial likelihood for the constrained fit with constraint s, we construct the GCV-style statistic

$$\mathrm{GCV}(s) = \frac{1}{N} \frac{-\ell_s}{N[1 - p(s)/N]^2}. \tag{6}$$

Intuitively, the GCV criterion inflates the negative log partial likelihood by a factor that involves $p(s)$, the effective number of parameters. Larger values of $p(s)$ cause more inflation (penalization) of the negative log partial likelihood. See Wahba[6] for details of generalized cross-validation; one could also use an Akaike-style criterion, as in Akaike.[7] Figure 2 shows the GCV plot for the lung cancer example, as a function of the standardized constraint parameter. The minimum occurs near $u = 0{\cdot}45$.

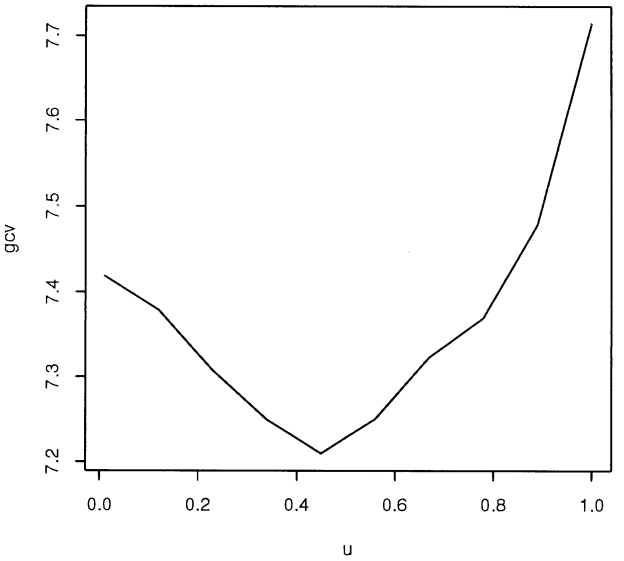

Figure 2. GCV plot for lung cancer example. The generalized cross-validation score is plotted against the standardized constraint parameter $u = s/\sum|\hat{\beta}_j^o|$

5. A SIMULATION STUDY

5.1. A few large effects

In this example we simulated 50 datasets each with 50 observations, from the exponential hazard model

$$\lambda(t|\mathbf{x}) = \exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$$

where $\boldsymbol{\beta} = (-0{\cdot}35, -0{\cdot}35, 0, 0, 0, -0{\cdot}35, 0, 0, 0)^{\mathrm{T}}$. The x_i were each marginally standard normal, and the correlation between x_i and x_j was $\rho^{|i-j|}$ with $\rho = 0{\cdot}5$. This gave moderate to strong effects for the three regressors with non-zero coefficients.

Letting $\sum$ be the population covariance matrix of the regressors, Table II shows the median of the mean squared errors $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}} \sum(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ over 50 simulations from this model. As before, the stepwise method refers to backward–forward stepwise selection as implemented in Scott Emerson's S language function 'coxrgrss' with the default P-values to enter and remove of 0·05 and 0·10, respectively.

The lasso clearly outperforms stepwise selection, and picks approximately the correct number of zero coefficients.

Figure 3 shows box plots of the coefficients from each of the three methods. The lasso does a better job of isolating the non-zero coefficients and shrinks the others substantially.

5.2. Many small effects

Here we used the same model as in the previous section, but with $\beta_j = 0{\cdot}1\ \forall_j$. This resulted in an occasional effect significant at the 0·05 level, as measured by the full Cox model fit. Table III shows the results. The lasso outperforms the full and stepwise models by shrinking the coefficients almost all of the way to zero.

Table II. Results for example 3 (a few large effects). Mean squared errors (MSE) over 50 simulations

Method	Median MSE (standard errors)	Average numbers of zero coefficients
Null	0·44 (—)	9·0
Full model	0·82 (0·13)	0·0
Stepwise	0·63 (0·12)	5·6
Lasso	0·26 (0·07)	6·7

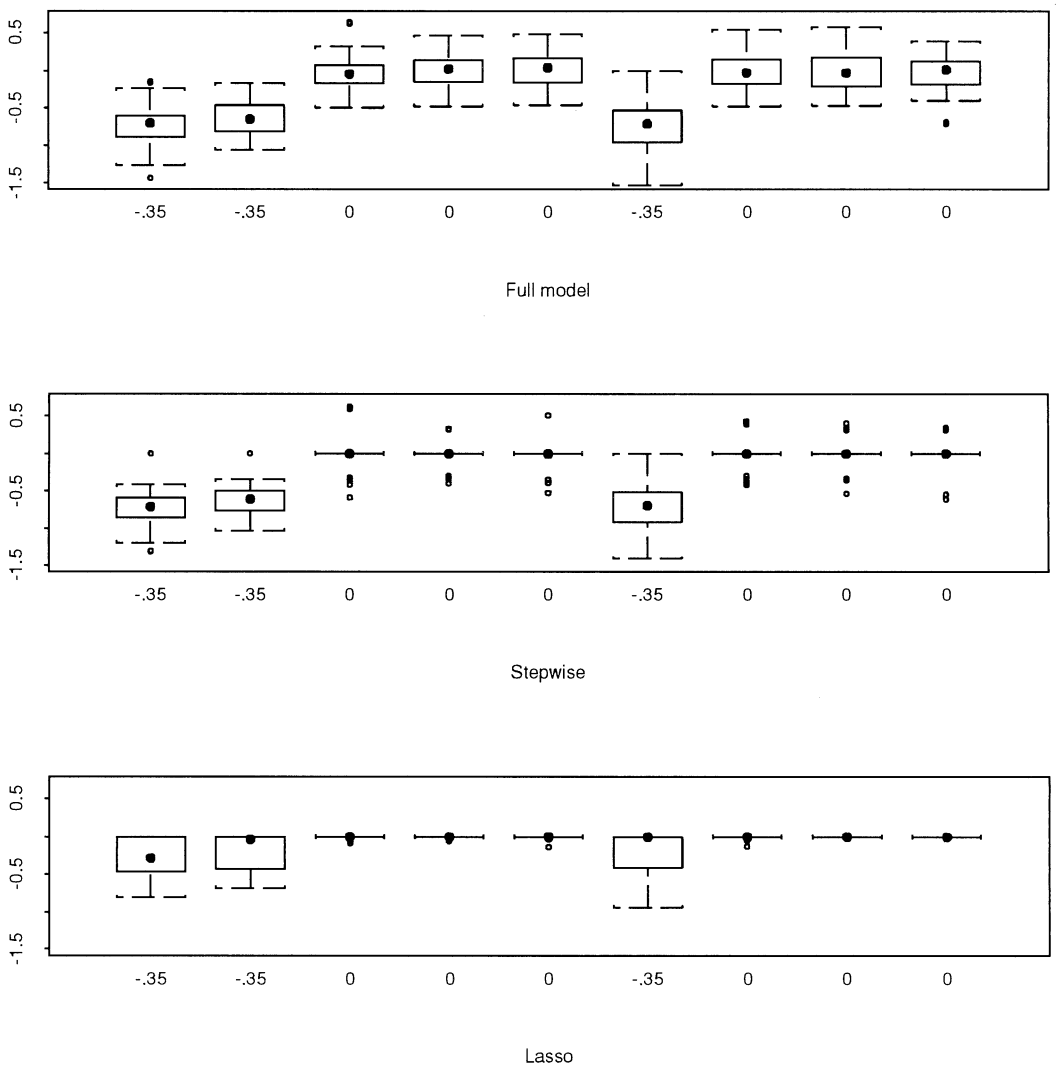

Figure 3. Box plot of coefficients from first simulation study (a few large coefficients)

Table III. Results for second simulation study (many small effects). Mean squared errors (MSE) over 50 simulations

Method	Median MSE (standard errors)	Average numbers of zero coefficients
Null	0·15 (—)	9·0
Full model	0·57 (0·04)	0·0
Stepwise	0·53 (0·04)	5·5
Lasso	0·15 (0·00)	7·8

Table IV. Estimated and actual standard errors for simulated example

Variable	$u = 0{\cdot}7$			$u = 0{\cdot}3$		
	Mean coefficient	Mean $\widehat{\text{SE}}$	Actual SE	Mean coefficient	Mean $\widehat{\text{SE}}$	Actual SE
1	− 0·55	0·17	0·19	− 0·30	0·13	0·16
2	− 0·58	0·18	0·24	− 0·35	0·15	0·17
3	− 0·01	0·09	0·16	− 0·02	0·02	0·06
4	0·01	0·07	0·10	0·00	0·00	0·01
5	0·00	0·07	0·13	− 0·01	0·01	0·05
6	− 0·50	0·16	0·20	− 0·23	0·12	0·12
7	− 0·05	0·09	0·15	− 0·01	0·01	0·03
8	0·00	0·09	0·15	0·00	0·01	0·02
9	− 0·01	0·06	0·11	0·00	0·00	0·01

6. STANDARD ERRORS

We can use approximation (5) to yield an approximate method for obtaining standard errors for the lasso estimates. In the notation of equation (5), we can show using standard partial likelihood theory that the variance of $\mathbf{z}$ is approximately $\mathbf{D}^{-1}$. Letting $\mathbf{M}$ denote the matrix that multiplies $\mathbf{z}$ in equation (5), then the variance of $\hat{\beta} = \mathbf{Mz}$ is approximately $\mathbf{MD}^{-1}\mathbf{M}^{\mathrm{T}}$. Hence we can obtain the approximate standard errors of $\hat{\beta}$ from the square root of the diagonal of $\mathbf{MD}^{-1}\mathbf{M}^{\mathrm{T}}$. To investigate the accuracy of this procedure, I simulated 50 datasets from the example of Section 5.1. The regressor variables were generated once and fixed for the 50 simulations. Table IV shows the results.

In the left part of the table the standardized constraint parameter was fixed at 0·7; in the right half, it was 0·3. In each half, the first column shows the mean coefficient over the 50 simulations, the second column gives the mean of the standard error estimate, and the third column shows the actual standard error of the coefficients over the 50 simulations. Note that on the average the standard error formula gives a reasonable estimate for the large effects (variables 1, 2, 6). For the small effects, it tends to underestimate the standard error, but not so much (except in one instance) as to affect that the perceived significance of the variable. Thus the approximate standard errors appear reliable, except when the estimated coefficient itself is very small. Note also that the sampling distribution of the estimates tends to be skewed, especially for small values of the constraint parameter (see Figure 3), so this will degrade the accuracy of normal confidence limits.

7. DISCUSSION

The lasso technique for variable selection in the Cox model seems a worthy competitor to stepwise selection. It is less variable than the stepwise approach and still yields interpretable models.

In practice the lasso should be used in conjunction with other model building tools. In particular, in our examples we have assumed that linearity is reasonable for all of the predictors. This may not be a good assumption and it should be checked in a real data analysis. Similarly, a referee pointed out that the proportional hazards assumption is unreasonable of cell type and Karnofsky score in the first example, and this should be checked; I have not attempted to provide thorough analyses in my examples.

The lasso method requires initial standardization of the regressors, so that the penalization scheme is fair to all regressors. For categorical regressors, one codes the regressor with dummy variables and then standardizes the dummy variables. As pointed out by a referee, however, the relative scaling between continuous and categorical variables in this scheme can be somewhat arbitrary.

In some problems there might be effects that the analyst does not want to shrink at all. Such an example is a treatment factor known to be effective *a priori*, with interest lying in more accurate adjustment for other covariates. In such an instance, one simply omits the corresponding β_j from the constraint, and one solves the problem by a simple modification of the optimization procedure discussed earlier.

There are some other recently proposed approaches to model selection. The Bayesian approach has been developed by Michell and Beauchamp[8] and George and McCulloch.[9] They use a hierarchical Bayes model; the second authors apply the Gibbs sampler to simulate a large collection of subset models from the posterior distribution. Sauerbrei and Schumacher[10] propose a bootstrap-based approach to model selection.

There are some ways to extend this work. This paper has focused on fixed covariates, but one can incorporate time-dependent covariates without any new difficulty. Adaptation of the lasso technique to the matched case control models is also fairly straightforward.

ACKNOWLEDGEMENTS

I wish to thank Michael LeBlanc for providing Fleming and Harrington's liver data, and Martin Schumacher and two referees for helpful comments. This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

1. Cox, D. 'Regression models and life tables (with discussion)', *Journal of the Royal Statistical Society, Series B*, **74**, 187–220 (1972).
2. Tibshirani, R. 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society, Series B*, (1995).
3. Hastie, T. and Tibshirani, R. *Generalized Additive Models*, Chapman and Hall, 1990.
4. Kalbfleisch, J. and Prentice, R. *The Statistical Analysis of Failure Time Data*, Wiley, New York, 1980.
5. Schwarz, G. 'Estimating the dimension of a model', *Annals of Statistics*, **6**, 461–464 (1978).
6. Wahba 'Spline bases, regularization, and generalized cross-validation for solving approximation problems with large quantities of noisy data', *in* Proceedings of the International Conference on Approximation theory in honour of George Lorenz, Academic Press, Austin, Texas, 1980.

7. Akaike, H. 'Information theory and an extension of the maximum likelihood principle', *in* Second International Symposium on Information Theory, 1973 pp. 267–281.
8. Mitchell, T. and Beauchamp, J. 'Bayesian variable selection in linear regression', *Journal of the American Statistical Association*, **83**, 1023–1036 (1988).
9. George, E. and McCulloch, R. 'Variable selection via Gibbs sampling', *Journal of the American Statistical Association*, **88**, 884–889 (1993).
10. Sauerbrei, W. and Schumacher, M. 'A bootstrap resampling procedure for model building: application to the cox regression model', *Statistics in Medicine*, **11**, 2093–2109 (1992).