Variable Selection for Survival Models

Wenbin Lu

Department of Statistics
North Carolina State University
lu@stat.ncsu.edu

March 24, 2016

Table of contents

Background and motivation

A review of semi-parametric survival models

- Cox's proportional hazards (PH) model (Cox, 1972):

$$\lambda(t|Z) = \lambda_0(t)\exp(\beta_0' Z)$$

- Proportional odds (PO) model (Pettitt, 1982, 1984; Bennett, 1983):

$$\{1 - S(t|Z)\}/S(t|Z) = [\{1 - S_0(t)\}/S_0(t)]\exp(\beta_0' Z)$$

- Linear transformation (LT) models (Clayton and Cuzick, 1985; Cheng, Wei and Ying, 1995):

$$H_0(T) = -\beta_0' Z + \epsilon$$

Variable selection problems for censored data

- Write the regression coefficients $\beta_0 = (\beta_{01}, \cdots, \beta_{0p})'$.
- Index set for important variables: $I = \{1 \leq j \leq p : \beta_{j0} \neq 0\}$
- Index set for unimportant variables: $U = \{1 \leq j \leq p : \beta_{j0} = 0\}$
- Assume $|I| = p_0 < p$. Write $\beta_0 = (\beta_{I0}', \mathbf{0}')'$.
- Define $\tilde{T} = \min(T, C)$ and $\delta = I(T \leq C)$. Given the observed data $(\tilde{T}_i, \delta_i, Z_i)$, $i = 1, \cdots, n$, the main goals of a variable selection procedure are:
  - to identify I and U correctly;
  - to provide good estimators for β_{I0}.

Oracle properties

An ideal variable selection procedure should asymptotically satisfy:

- produce parsimonious models automatically (with probability one)

$$\hat{\beta}_j \neq 0 \quad \text{for} \quad j \in I$$

$$\hat{\beta}_j = 0 \quad \text{for} \quad j \in U;$$

- achieve the optimal estimation rate

$$\sqrt{n}(\hat{\beta}_I - \beta_{I0}) \to_d N(0, \Sigma_{I0}),$$

where Σ_{I0} is the covariance matrix knowing the true model.

Oracle procedure performs as well as if the correct true model were known.

Existing variable selection methods for censored data

- Best subset selection and stepwise selection
- Asymptotic testing procedures, such as score test and Wald test
- Bootstrap sampling procedures (Sauerbrei and Schumacher 1992)
- Bayesian variable selection (Faraggi and Simon 1998; Ibrahim, Chen and MacEachern 1999)
- Shrinkage methods (LASSO: Tibshirani 1997; SCAD: Fan and Li 2002; Adaptive-LASSO: Zhang and Lu 2007)

Penalized partial likelihood estimation for Cox's model

- Log partial likelihood (Cox 1975):

$$l_n(\beta) \equiv \sum_{i=1}^{n} \delta_i \{\beta' Z_i - \log[\sum_{j=1}^{n} I(\tilde{T}_j \geq \tilde{T}_i) \exp(\beta' Z_j)]\}.$$

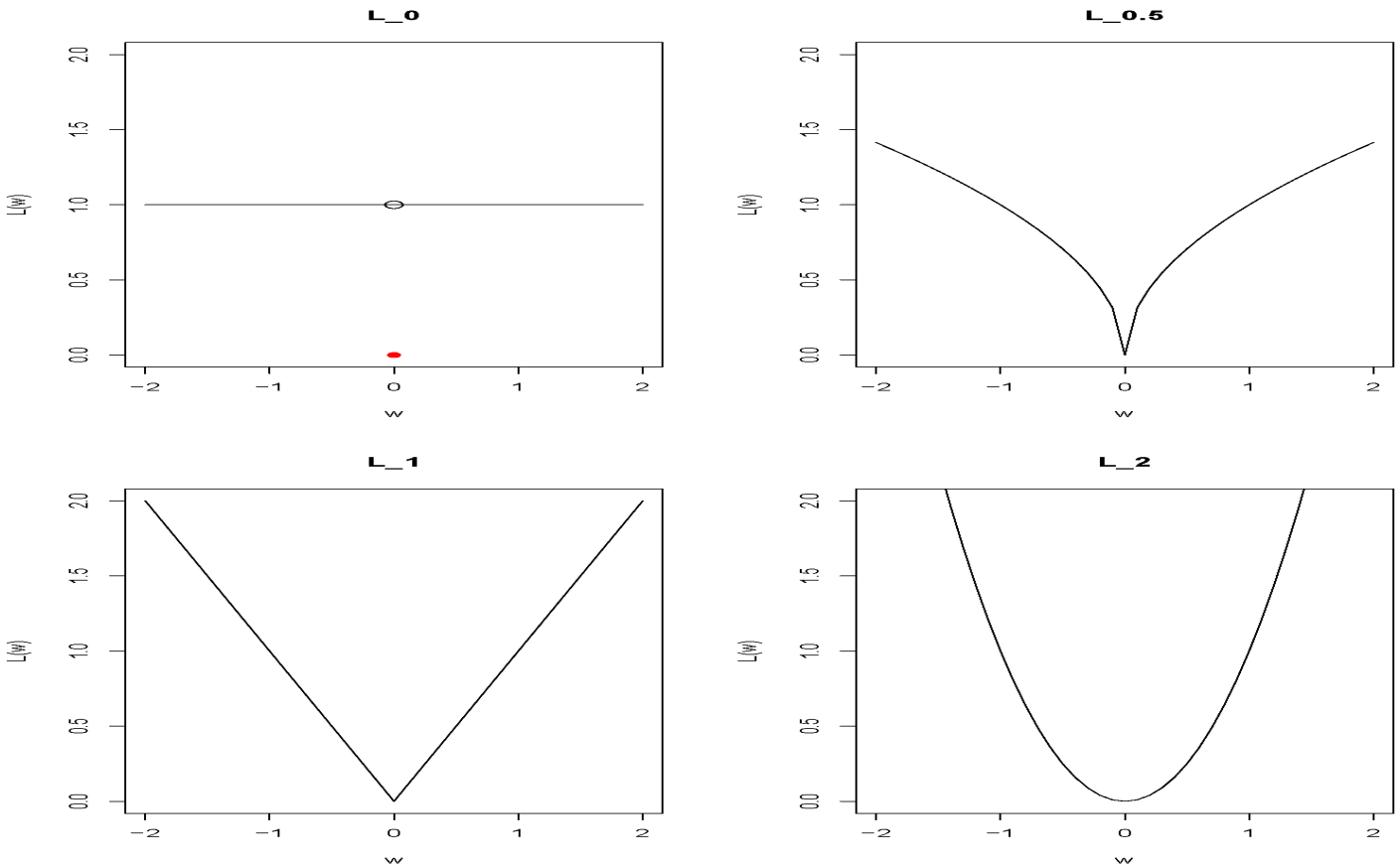
- The penalized log partial likelihood estimation

$$\min_{\beta} -\frac{1}{n} l_n(\beta) + \sum_{j=1}^{p} J_\lambda(\beta_j).$$

Choices of penalty function

- Ridge regression (Hoerl and Kennard, 1970): $J_\lambda(\beta_j) = \lambda\beta_j^2$.
- Bridge regression (Frank and Friedman, 1993): $J_\lambda(\beta_j) = \lambda|\beta_j|^q, \quad q \geq 0$.
  - If $q = 0$, known as entropy penalty (Donoho and Johnstone, 1998).
  - If $q = 1$, known as LASSO (Tibshirani, 1996).
  - For $q \leq 1$, it tends to shrink small $|\beta|$'s to exactly zero.
  - J_λ is not convex for $q < 1$ while solutions are not sparse for $q > 1$.

L_q penalty functions



Properties of LASSO estimators

- The lasso has shown good performance in practice.
- In general, the lasso may not be consistent for variable selection. Under linear regression model settings:
  - If $\lambda = O(\sqrt{n})$, the lasso is root-n consistent. (Knight and Fu, 2002)
  - If $\lambda = O(\sqrt{n})$, Zou (2006) showed that

$$\lim \sup_{n \to \infty} P(\hat{I}_n = I) \leq c < 1,$$

where $\hat{I}_n$ is the index set of variables selected by the lasso.

The lasso estimator is not *oracle*.

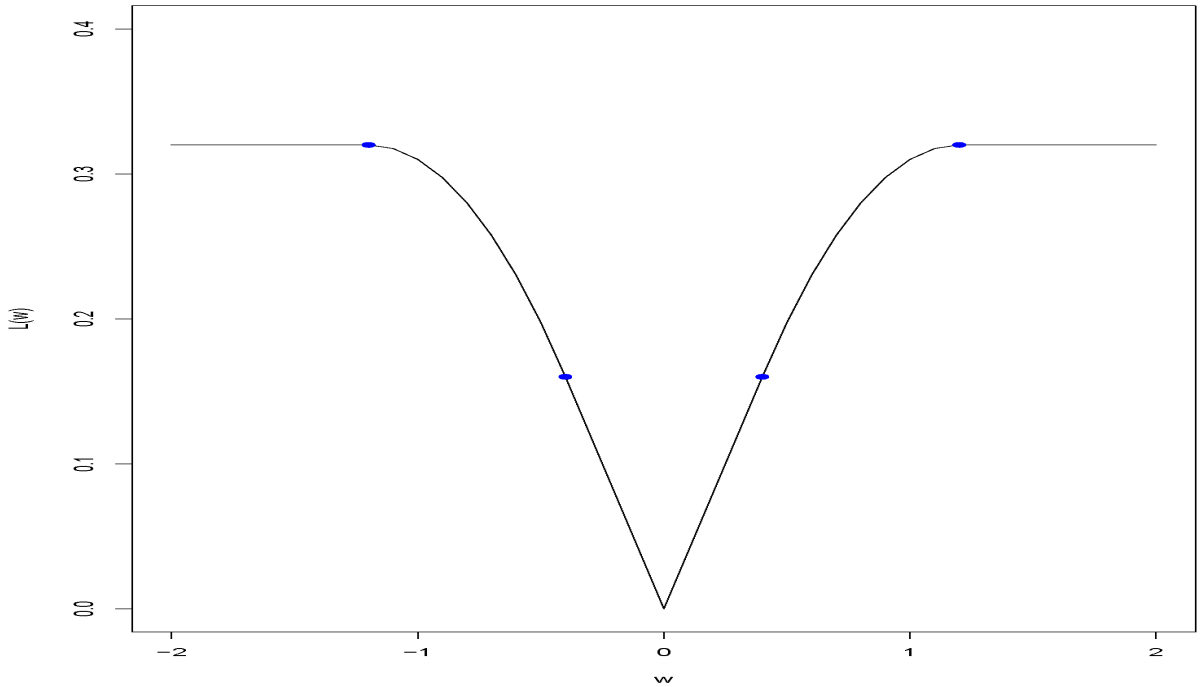
Smoothly clipped absolute deviation (SCAD) penalty

Fan and Li (2001) suggested to use

$$J_\lambda(w) = \begin{cases} \lambda|w| & \text{if} \quad |w| \le \lambda, \\ -\frac{(|w|^2 - 2a\lambda|w| + \lambda^2)}{2(a-1)} & \text{if} \quad \lambda < |w| \le a\lambda, \\ \frac{(a+1)\lambda^2}{2} & \text{if} \quad |w| > a\lambda, \end{cases}$$

where $a > 2$ and $\lambda > 0$ are tuning parameters.

An example of (SCAD) Penalty



where $\lambda = 0.4$ and $a = 3$.

Properties of SCAD penalty

- A quadratic spline function with two knots at λ and $a\lambda$.
- Except being singular at the origin, the function $p_\lambda(|w|)$ has a continuous first-order derivative.
- The SCAD is *oracle* when λ is properly tuned (Fan and Li, 2001)
- But it is not convex! (challenging to implement)

Adaptive LASSO estimation

We solve (Zhang and Lu, 2007)

$$\min_{\beta} -\frac{1}{n} l_n(\beta) + \lambda \sum_{j=1}^{p} |\beta_j| w_j,$$

where $\boldsymbol{w} = (w_1, \cdots, w_p)'$ are the data-dependent weights.
Key Motivations:

- Large penalties are imposed on unimportant covariate effects, while small penalties for important ones. (Protect important covariates more)
- Let the data choose w_j's adaptively.

Discussions and extensions

- The choice of w_j's
  - The appropriate values of w_j's will guarantee the optimality of the adaptive-LASSO solution.
  - We propose using $w_j = 1/|\tilde{\beta}_j|$, where $\tilde{\beta} = (\tilde{\beta}_1, \cdots, \tilde{\beta}_p)'$ are maximum partial likelihood estimators.
  - Any root-n consistent estimates of β's can be used, and $\tilde{\beta}$ is just a convenient choice.
  - It is closely related to the L_0 penalty $\sum_{j=1}^{p} I(|\beta_j| \neq 0)$ (Donoho & Johnstone 1998; Antoniadis & Fan 2001).
- Extensions
  - PO model: penalized marginal likelihood (PML) (Lu and Zhang, 2007);
  - Linear transformation model: penalized estimating equation (PEE) (Zhang, Lu and Wang, 2010)

Simulation studies

- We consider PH and PO models.
- We choose $\beta = (-1, -0.9, 0, 0, 0, -0.8, 0, 0, 0)'$, and the nine covariates $Z = (Z_1, ..., Z_9)$ are marginally standard normal with the pairwise correlation $\text{corr}(Z_j, Z_k) = \rho^{|j-k|}$ with $\rho = 0.5$.
- Censoring times are from uniform (0,c): 25% and 40% censoring rates
- Sample sizes n = 100, 200, simulation replications $M = 500$.
- We compare the PEE (Zhang, Lu and Wang, 2010), PPL (Zhang and Lu, 2007), PML (Lu and Zhang, 2007) estimates.

Simulation results for PH model

Table 1. Mean squared error and model selection results

n	Censored	Method	Average MSE	Model Size	Number of zero coefficients	
				oracle (3)	correct (6)	incorrect (0)
100	25%	EE	0.244 (0.161)	9	0 (0)	0 (0)
		PEE	0.122 (0.119)	3.610 (0.920)	5.390 (0.920)	0.000 (0.000)
		PPL	0.130 (0.121)	3.136 (0.412)	5.858 (0.403)	0.006 (0.077)
	40%	EE	0.277 (0.186)	9	0 (0)	0 (0)
		PEE	0.143 (0.133)	3.620 (0.885)	5.380 (0.885)	0.000 (0.000)
		PPL	0.177 (0.161)	3.150 (0.456)	5.836 (0.435)	0.014 (0.118)
200	25%	EE	0.087 (0.052)	9	0 (0)	0 (0)
		PEE	0.051 (0.040)	3.250 (0.557)	5.750 (0.557)	0.000 (0.000)
		PPL	0.053 (0.050)	3.034 (0.181)	5.966 (0.181)	0.000 (0.000)
	40%	EE	0.110 (0.066)	9	0 (0)	0 (0)
		PEE	0.063 (0.049)	3.280 (0.604)	5.720 (0.604)	0.000 (0.000)
		PPL	0.062 (0.055)	3.048 (0.214)	5.952 (0.214)	0.000 (0.000)

Simulation results for PO model

Table 2. Mean squared error and model selection results

n	Censored	Method	Average MSE	Model Size	Number of zero coefficients	
				oracle (3)	correct (6)	incorrect (0)
100	25%	EE	0.481 (0.262)	9	0 (0)	0 (0)
		PEE	0.377 (0.303)	3.600 (0.932)	5.230 (0.874)	0.170 (0.403)
		PML	0.436 (0.419)	2.898 (0.684)	5.856 (0.389)	0.246 (0.539)
	40%	EE	0.575 (0.347)	9	0 (0)	0 (0)
		PEE	0.385 (0.314)	3.490 (0.916)	5.360 (0.811)	0.150 (0.386)
		PML	0.493 (0.484)	2.834 (0.735)	5.844 (0.400)	0.322 (0.599)
200	25%	EE	0.213 (0.109)	9	0 (0)	0 (0)
		PEE	0.122 (0.085)	3.340 (0.670)	5.660 (0.670)	0.000 (0.000)
		PML	0.231 (0.120)	3.026 (0.193)	5.968 (0.176)	0.006 (0.077)
	40%	EE	0.258 (0.168)	9	0 (0)	0 (0)
		PEE	0.132 (0.086)	3.310 (0.598)	5.690 (0.598)	0.000 (0.000)
		PML	0.218 (0.142)	3.030 (0.239)	5.952 (0.214)	0.018 (0.133)

Primary biliary cirrhosis data

- Data gathered in the Mayo Clinic trial in primary biliary cirrhosis of liver conducted between 1974 and 1984 (Therneau and Grambsch 2000).
- 312 eligible subjects with 125 deaths
- 17 predictors: 10 continuous and 7 discrete.
- Goal: to study the dependence of survival times on 17 covariates.
- Zhang and Lu (2007) studied variable selection for this data in the PH model using the penalized partial likelihood method with the adaptive Lasso penalty.

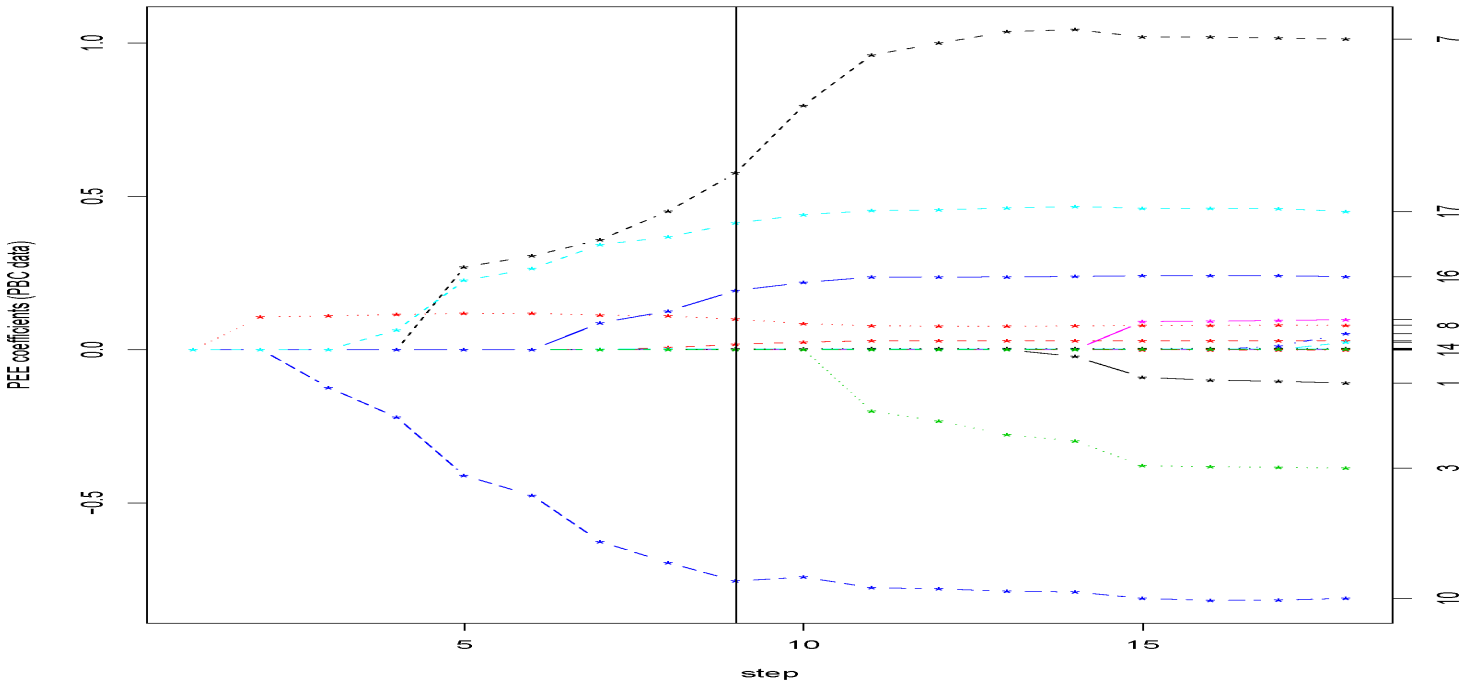
Analysis of PBC data

Table 4. Estimation and variable selection for PBC data with the PH model.

Covariate	EE	PEE	PPL
trt	-0.109 (0.234)	0 (-)	0 (-)
age	0.029 (0.012)	0.017 (0.007)	0.019 (0.010)
sex	-0.386 (0.346)	0 (-)	0 (-)
asc	0.053 (0.469)	0 (0)	0 (-)
hep	0.024 (0.263)	0 (-)	0 (-)
spid	0.098 (0.279)	0 (-)	0 (-)
oed	1.013 (0.486)	0.576 (0.241)	0.671 (0.377)
bil	0.079 (0.024)	0.099 (0.018)	0.095 (0.020)
chol	0.001 (0.000)	0 (-)	0 (-)
alb	-0.811 (0.286)	-0.755 (0.211)	-0.612 (0.280)
cop	0.003 (0.001)	0.003 (0.001)	0.002 (0.001)
alk	0.000 (0.000)	0 (-)	0 (-)
sgot	0.004 (0.002)	0.002 (0.001)	0.002 (0.001)
trig	-0.001 (0.001)	0 (-)	0 (-)
plat	0.001 (0.001)	0 (-)	0 (-)
prot	0.238 (0.103)	0.193 (0.066)	0.103 (0.108)
stage	0.450 (0.171)	0.413 (0.121)	0.367 (0.142)

Solution path for the PEE estimates

- For PBC data using PH model



Lung cancer data

- Data is from the Veteran's Administration lung cancer trial (Kalbfleish and Prentice 2002).
- 137 males with advanced inoperable lung cancer were randomized to either a standard treatment or chemotherapy
- There are six covariates: Treatment (1=standard, 2=test), Cell type (1=squamous, 2=small cell, 3=adeno, 4=large), Karnofsky score, Months from Diagnosis, Age, and Prior therapy (0=no, 10=yes).
- Lu and Zhang (2007) studied variable selection for this data in the PO model using the penalized marginal likelihood method with the adaptive Lasso penalty.

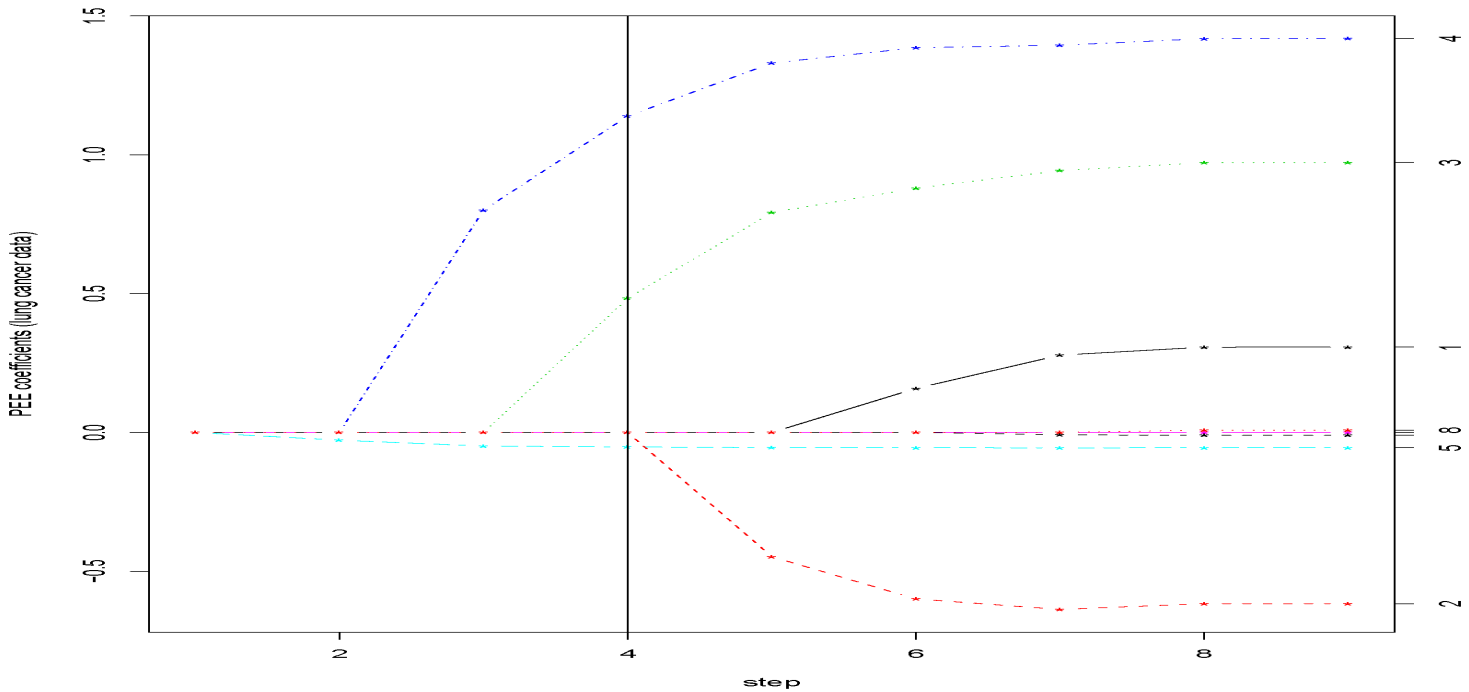
Analysis of lung cancer data

Table 5. Estimation and variable selection results for lung cancer data with the PO model.

Covariate	EE	PEE	PML
Treatment	0.307 (0.317)	0 (-)	0 (-)
squamous vs large	-0.617 (0.482)	0 (-)	0 (-)
small vs large	0.972 (0.473)	0.483 (0.197)	0.706 (0.356)
adeno vs large	1.418 (0.371)	1.139 (0.261)	0.841 (0.397)
Karnofsky	-0.055 (0.009)	-0.052 (0.008)	-0.053 (0.008)
Months from Diagnosis	0.000 (0.015)	0 (-)	0 (-)
Age	-0.010 (0.017)	0 (-)	0 (-)
Prior therapy	0.008 (0.040)	0 (-)	0 (-)

Solution path for the PEE estimates

- For lung cancer data using PO model



Softwares and future directions

- Softwares
  - My web link:
    `http://www4.stat.ncsu.edu/~lu/programcodes.html`
  - R package: glmnet (for Cox model)
- References
  - Zhang, H. H. and Lu, W. (2007). Adaptive-LASSO for Cox's Proportional Hazards Model. Biometrika, 94, 1-13.
  - Lu, W. and Zhang, H. H. (2007). Variable Selection for Proportional Odds Model. Statistics in Medicine, 26, 3771-3781.
  - Zhang, H. H., Lu, W. and Wang, H. (2010) On sparse estimation for semiparametric linear transformation models. Journal of Multivariate Analysis, 101, 1594-1606.