MATH 686 Project

Daniel Krasnov

Introduction

First, the SCAD and LASSO methods are introduced for variable selection in the Cox's Proportional Hazards Model. Then results from (Fan and Li 2002) are replicated which compares these variable selection methods with AIC and BIC best subset selection. Finally a dataset is introduced to which SCAD and LASSO will be applied for my MATH 686 project.

Background

INTRODUCE COX'S MODEL HERE

The following is taken from (Fan and Li 2002). Consider independent samples $(\mathbf{x}_i, Y_i)$ with conditional density $f_i(y_i; \mathbf{X}_i^T, \ )$. Let $\ell_i = \log f_i$. Then a general form of penalized likelihood is given by

$$\sum_{i=1}^{n} \ell_i y_i; \mathbf{x}_i^T \ \ - n \sum_{j=1}^{d} p_\lambda(|\beta_j|)$$

where d is the dimension of , $p_\lambda(.)$ is some penalty function and $|lambda$ is a tuning parameter. Selecting a function $p_\lambda(.)$ amounts to selecting a variable selection method for COx's Proportional Hazards model. In this study we will consider two penalties: LASSO and SCAD,

$$\begin{aligned} p_\lambda(|\theta|) &= \lambda|\theta| && \text{LASSO} \\ p_\lambda(\theta) &= I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda) && \text{SCAD} \end{aligned}$$

where $a > 2$ and $\theta > 0$. In general, a value of $a = 3.7$ is used and we adopt this for our study.

SCAD and LASSO Simulation

In this section model performance is compared on simulated data for the LASSO, SCAD, and AIC and BIC best subset regression parameter selection techniques. Model performance is assessed through the Relative Model Errors (RME)

$$\mathbb{E}\left\{\exp(-\mathbf{X}^T\hat{\beta}) - \exp(-\mathbf{X}^T\beta_0)\right\}^2.$$

We simulate 100 datasets with $n = 75$ and $n = 100$ observations from the exponential hazard model

$$h(t|\mathbf{x}) = \exp(\mathbf{x}^T\ )$$

- $\beta = (0.8, 0, 0, 1, 0, 0, 0.6, 0)^T$,
- x_i are marginally standard normal with correlation $\rho = 0.5$,
- Censoring times are exponentially distributed with mean $U\exp(\mathbf{x}^{\mathbf{T}}\ _{\mathbf{0}})$, $U \sim \text{Uniform}(1, 3)$.

Our simulations yielded the following results:

Method	MRME(%)	Aver. no. cor. 0 coeff.	Aver. no. incor. 0 coeff.
n=75			
SCAD	73.53	4.07	0.05
LASSO	51.23	2.85	0.01
AIC	73.95	4.23	0.02
BIC	56.21	4.78	0.06
n=100			
SCAD	55.21	4.21	0
LASSO	39.55	2.87	0
AIC	74.03	4.13	0.01
BIC	55.31	4.71	0.01

Method	Beta1 SD	Beta4 SD	Beta7 SD
n=75			
SCAD	0.25	0.24	0.29
LASSO	0.2	0.23	0.21
AIC/BIC	0.21	0.24	0.22
n=100			
SCAD	0.17	0.18	0.18
LASSO	0.17	0.18	0.14

Method	Beta1 SD	Beta4 SD	Beta7 SD
AIC/BIC	0.18	0.18	0.22

Dataset

Data

```
library(survival)
data(pbc)
kable(head(pbc))
```

id	time	status	trt	age	sex	ascites	hepato	spiders	edema	bili	chol	albumin	copper	alk.phos	ast	trig	platelet	protime	stage
1	400	2	1	58.76523	f	1	1	1	1.0	14.5	261	2.60	156	1718.0	137.95	172	190	12.2	4
2	4500	0	1	56.44627	f	0	1	1	0.0	1.1	302	4.14	54	7394.8	113.52	88	221	10.6	3
3	1012	2	1	70.07255	m	0	0	0	0.5	1.4	176	3.48	210	516.0	96.10	55	151	12.0	4
4	1925	2	1	54.74059	f	0	1	1	0.5	1.8	244	2.54	64	6121.8	60.63	92	183	10.3	4
5	1504	1	2	38.10541	f	0	1	1	0.0	3.4	279	3.53	143	671.0	113.15	72	136	10.9	3
6	2503	2	2	66.25873	f	0	1	0	0.0	0.8	248	3.98	50	944.0	93.00	63	NA	11.0	3

Here is the description of the dataset from the R help file:

Primary biliary cholangitis is an autoimmune disease leading to destruction of the small bile ducts in the liver. Progression is slow but inexhortable, eventually leading to cirrhosis and liver decompensation. The condition has been recognised since at least 1851 and was named "primary biliary cirrhosis" in 1949. Because cirrhosis is a feature only of advanced disease, a change of its name to "primary biliary cholangitis" was proposed by patient advocacy groups in 2014.

This data is from the Mayo Clinic trial in PBC conducted between 1974 and 1984. A total of 424 PBC patients, referred to Mayo Clinic during that ten-year interval, met eligibility criteria for the randomized placebo controlled trial of the drug D-penicillamine. The first 312 cases in the data set participated in the randomized trial and contain largely complete data. The 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 the data here are on an additional 106 cases as well as the 312 randomized participants.

The variables are as follows:

Variable	Description
age	in years
albumin	serum albumin (g/dl)
alk.phos	alkaline phosphatase (U/liter)
ascites	presence of ascites
ast	aspartate aminotransferase, once called SGOT (U/ml)
bili	serum bilirubin (mg/dl)
chol	serum cholesterol (mg/dl)
copper	urine copper (ug/day)
edema	0 no edema, 0.5 untreated or successfully treated, 1 edema despite diuretic therapy
hepato	presence of hepatomegaly or enlarged liver
id	case number
platelet	platelet count
protime	standardized blood clotting time
sex	m/f
spiders	blood vessel malformations in the skin
stage	histologic stage of disease (needs biopsy)
status	status at endpoint, 0/1/2 for censored, transplant, dead
time	number of days between registration and the earlier of death, transplantation, or study analysis in July, 1986
trt	1/2/NA for D-penicillamine, placebo, not randomized
trig	triglycerides (mg/dl)

References

Fan, Jianqing, and Runze Li. 2002. "Variable Selection for Cox's Proportional Hazards Model and Frailty Model." *The Annals of Statistics* 30 (1): 74–99.