Corrections are marked yellow

patients were lost to followup soon after their initial clinic visit and excluded from our study. The rest of the 106 cases did not participate in the randomized trial but were followed for survival with some basic measurements recorded.

Since there was no treatment difference with respect to the survival distributions at the end of study, the study investigators combined the data from the two treatment groups to establish models for predicting survival. In this article, we utilized data on all 418 patients to establish a prediction model for the patient's survival given their baseline covariates. The average follow-up time of these 418 patients was 5.25 years. Like other studies, there were missing covariate values among the patients ranging from 2 patients missing prothrombin time (protime) to 136 patients missing triglyceride levels. For illustration, we imputed the missing values with their group sample median.

The outcome variable is the time to death ($time_i$). Censoring variable ($death_i$) for each case i has value 1 if the death date is available, or value 0 otherwise. The patient's baseline information consists of

- Demographic attributes: age in years, sex
- Clinical aspects: ascites (presence/absence), hepatomegaly (presence/absence), spiders (blood vessel malformations in the skin, presence/absence), edema (0 no edema and no diuretic therapy for edema, 0.5 edema untreated or successfully treated, 1 edema despite diuretic therapy)
- Biochemical aspects: serum bilirubin (mg/dl), albumin (g/dl), urine copper (μg/day), prothrombin time (standardised blood clotting time in seconds), platelet count n (number of platelets $\times$ 10^{-3} per mL^3), alkaline phosphotase (U/liter), ast (aspartate aminotransferase, once called SGOT (U/ml)), serum cholesterol (mg/dl), and triglyceride levels (mg/dl)
- Histologic stage of disease.

We applied logarithmic transformations to albumin, bilirubin, and protime in the process of model building, based on analyses of this dataset in Fleming and Harrington (1991).

To establish a prediction model, ideally one should have three similar but independent datasets, or split the dataset randomly into three subsets. Using the observations from the first subset, we fit the data with all model candidates of interest; using the data from the second piece, we evaluate those fitted models with intuitively interpretable, model-free criteria and choose a final model; and using the data from the third piece, we draw inferences about the selected model. In practice, if the sample size is not large, we may combine the first two steps with a cross-validation procedure.

We will use the PBC dataset to illustrate this model selection strategy in Section 13.6. First we review some classical algorithms for model selection and introduce some model-free, heuristically interpretable criteria for model evaluation.

13.3 Model building procedures and evaluation

Depending on the study question and subject matter knowledge, we may identify a set of potential explanatory variables which could be associated with the survival outcome in a Cox PH model, the hazard function at time t for an individual is:

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

where $\lambda_0(t)$ is an unknown baseline hazard function, $Z = (z_1, z_2, \ldots, z_p)'$ is the vector of explanatory variables of the individual, and $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ is a $p \times 1$ vector of

TABLE 13.1
Models derived from various classical model selection methods, using entire PBC dataset, hazard ratios $\exp(\hat{\beta})$ are presented.

Covariates	M1	M2	M3	M4	Lasso[@]		Ridge	
					AIC	BIC	AIC	BIC
logbili	2.334	2.372	2.279	2.688	2.213	2.137	2.016	1.651
edema	2.238	2.110	2.107		2.022	1.996	2.182	2.099
age	1.034	1.032	1.034		1.031	1.027	1.029	1.023
stage	1.386	1.394	1.412		1.366	1.326	1.369	1.284
lalb	0.120	0.119	0.128		0.168	0.180	0.166	0.199
lptime	8.164	7.267	8.004		6.535	5.513	7.715	6.898
ast			1.002		1.002	1.001	1.002	1.002
copper		1.002	1.001		1.001	1.001	1.001	1.002
ascites					1.320	1.291	1.407	1.498
trig		0.998	0.998		0.998	0.999	0.998	0.999
hepato					1.049		1.158	1.223
spiders							0.947	1.044
sex							1.057	1.063
chol							1.000	1.000
alk.phos							1.000	1.000
platelet							1.000	1.000

Note: All covariates were treated as continuous effects.

Model	Selection Method
M1	Several model building procedures using BIC as stopping criterion came up with this same model: a. Forward selection, BIC; b. Backward elimination, BIC; c. Stepwise, BIC
M2	Backward elimination, AIC
M3	a. Forward selection, AIC; b. Stepwise, AIC
M4	Best single variable model, logbili (log(bilirubin)) is the most significant variable ($p < .00001$)

the K-M survival curve, over the range from $[0, t_{max}]$, where t_{max} (= 12.5 years) is the maximum time for all K-M curves considered and serves as a common upper limit for the restricted mean calculation. The overall logrank test, and the logrank tests for the difference in survival distributions between any two risk categories all yield p-values < 0.00001. Both M1 and Ridge, BIC models produce similar results with little difference in C-Statistics, this further illustrates that M1 model is most preferable because it only takes 6 variables to achieve similar predictability.

TABLE 13.2
C-statistic of models using the full PBC dataset.

Model Selection Method	Model Size	C-Statistic
M4	1	0.748
M2	8	0.784
M1	6	0.790
M3	9	0.791
Lasso, AIC	11	0.794
Lasso, BIC	10	0.794
Ridge, AIC	16	0.796
Ridge, BIC	16	0.799

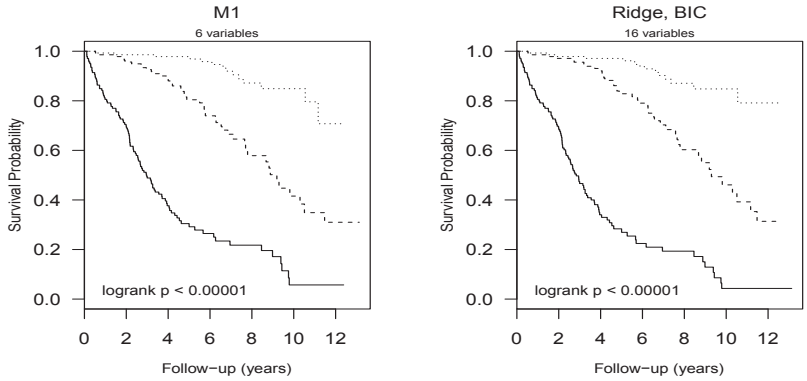

FIGURE 13.1
Kaplan-Meier curves of the survival time, stratified by tertiles of risk scores from two models: M1 - Six-variable model (left panel), and Ridge, BIC model (right panel).

TABLE 13.3
Summary statistics of the survival distributions by risk categories, scoring using the entire dataset.

Model Selection Method	Risk Categories	N Events/ Total	Restricted Mean (se) in years	Median (years)	(95% CI)
Stepwise, BIC	Low	14/140	11.31 (0.283)	NA	(NA, NA)
	Medium	45/139	8.66 (0.393)	9.19	(7.70, 11.47)
	High	102/139	4.21 (0.340)	2.97	(2.55, 3.71)
Ridge, BIC	Low	14/140	11.36 (0.278)	NA	(NA, NA)
	Medium	42/139	8.98 (0.381)	9.30	(7.79, NA)
	High	105/139	3.98 (0.320)	2.84	(2.44, 3.55)

13.5 Challenges and a proposal

The aforementioned process of using the same dataset for model building, selection, and inference has been utilized in practice. This conventional process has potential of self-serving