stepwise_cox_cx

Chuyuan XU

2025-12-12

Data Preparation

Data Wrangling (wenjie, zhaokun)

All rows with NA omitted for Stepwise and LASSO Cox model selection. Log-transformed `Bilirubin` is included in the Cox model. `Edama` derives a binary variable `Edema_bin` based on the presence or absence of Edema.

Stepwise Cox

```
## Start:  AIC=1080.53
## Surv(n_days, event) ~ drug + sex + ascites + hepatomegaly + spiders +
##     edema_bin + albumin + copper + alk_phos + sgot + platelets +
##     prothrombin + stage + age_years + log_bilirubin
##
##                 Df    AIC
## - platelets      1 1078.5
## - spiders        1 1078.6
## - hepatomegaly   1 1078.6
## - ascites        1 1078.7
## - drug           1 1078.8
## - sex            1 1079.0
## - alk_phos       1 1079.2
## - stage          3 1079.5
## - sgot           1 1080.4
## <none>             1080.5
## - copper         1 1080.7
## - edema_bin      1 1082.5
## - age_years      1 1085.3
## - prothrombin    1 1086.1
## - albumin        1 1086.5
## - log_bilirubin  1 1107.1
##
## Step:  AIC=1078.53
## Surv(n_days, event) ~ drug + sex + ascites + hepatomegaly + spiders +
##     edema_bin + albumin + copper + alk_phos + sgot + prothrombin +
##     stage + age_years + log_bilirubin
##
##                 Df    AIC
## - spiders        1 1076.6
## - hepatomegaly   1 1076.6
## - ascites        1 1076.7
## - drug           1 1076.8
```

```
## - sex            1 1077.0
## - alk_phos       1 1077.2
## - stage          3 1077.6
## <none>             1078.5
## - sgot           1 1078.6
## - copper         1 1078.7
## + platelets      1 1080.5
## - edema_bin      1 1080.8
## - age_years      1 1083.3
## - prothrombin    1 1084.2
## - albumin        1 1084.5
## - log_bilirubin  1 1106.1
##
## Step:  AIC=1076.56
## Surv(n_days, event) ~ drug + sex + ascites + hepatomegaly + edema_bin +
##     albumin + copper + alk_phos + sgot + prothrombin + stage +
##     age_years + log_bilirubin
##
##                 Df    AIC
## - hepatomegaly   1 1074.6
## - ascites        1 1074.7
## - drug           1 1074.9
## - sex            1 1075.1
## - alk_phos       1 1075.2
## - stage          3 1075.6
## <none>             1076.6
## - copper         1 1076.7
## - sgot           1 1076.7
## + spiders        1 1078.5
## + platelets      1 1078.6
## - edema_bin      1 1078.8
## - age_years      1 1081.5
## - prothrombin    1 1082.2
## - albumin        1 1082.5
## - log_bilirubin  1 1104.9
##
## Step:  AIC=1074.6
## Surv(n_days, event) ~ drug + sex + ascites + edema_bin + albumin +
##     copper + alk_phos + sgot + prothrombin + stage + age_years +
##     log_bilirubin
##
##                 Df    AIC
## - ascites        1 1072.7
## - drug           1 1072.9
## - sex            1 1073.2
## - alk_phos       1 1073.3
## - stage          3 1074.5
## <none>             1074.6
## - copper         1 1074.7
## - sgot           1 1074.7
## + hepatomegaly   1 1076.6
## + spiders        1 1076.6
## + platelets      1 1076.6
## - edema_bin      1 1076.9
```

```
## - age_years      1 1079.5
## - prothrombin    1 1080.3
## - albumin        1 1080.7
## - log_bilirubin  1 1105.6
##
## Step:  AIC=1072.74
## Surv(n_days, event) ~ drug + sex + edema_bin + albumin + copper +
##     alk_phos + sgot + prothrombin + stage + age_years + log_bilirubin
##
##                 Df    AIC
## - drug           1 1071.0
## - sex            1 1071.4
## - alk_phos       1 1071.6
## - stage          3 1072.7
## <none>             1072.7
## - sgot           1 1072.8
## - copper         1 1073.4
## + ascites        1 1074.6
## + spiders        1 1074.7
## + hepatomegaly   1 1074.7
## + platelets      1 1074.7
## - edema_bin      1 1075.2
## - age_years      1 1077.8
## - prothrombin    1 1078.6
## - albumin        1 1080.9
## - log_bilirubin  1 1104.8
##
## Step:  AIC=1071.03
## Surv(n_days, event) ~ sex + edema_bin + albumin + copper + alk_phos +
##     sgot + prothrombin + stage + age_years + log_bilirubin
##
##                 Df    AIC
## - sex            1 1069.6
## - alk_phos       1 1069.9
## <none>             1071.0
## - stage          3 1071.1
## - sgot           1 1071.1
## - copper         1 1071.8
## + drug           1 1072.7
## + ascites        1 1072.9
## + spiders        1 1073.0
## + hepatomegaly   1 1073.0
## + platelets      1 1073.0
## - edema_bin      1 1073.3
## - age_years      1 1075.8
## - prothrombin    1 1077.1
## - albumin        1 1079.0
## - log_bilirubin  1 1102.8
##
## Step:  AIC=1069.63
## Surv(n_days, event) ~ edema_bin + albumin + copper + alk_phos +
##     sgot + prothrombin + stage + age_years + log_bilirubin
##
##                 Df    AIC
```

```
## - alk_phos       1 1068.6
## - stage          3 1069.5
## <none>             1069.6
## - sgot           1 1069.9
## + sex            1 1071.0
## + drug           1 1071.4
## - edema_bin      1 1071.5
## + ascites        1 1071.5
## + spiders        1 1071.5
## + hepatomegaly   1 1071.6
## + platelets      1 1071.6
## - copper         1 1072.1
## - prothrombin    1 1075.5
## - age_years      1 1076.6
## - albumin        1 1077.2
## - log_bilirubin  1 1101.0
##
## Step:  AIC=1068.62
## Surv(n_days, event) ~ edema_bin + albumin + copper + sgot + prothrombin +
##     stage + age_years + log_bilirubin
##
##                 Df    AIC
## <none>             1068.6
## - stage          3 1068.9
## - sgot           1 1069.2
## + alk_phos       1 1069.6
## + sex            1 1069.9
## + drug           1 1070.3
## + ascites        1 1070.3
## - copper         1 1070.3
## + platelets      1 1070.5
## + hepatomegaly   1 1070.6
## + spiders        1 1070.6
## - edema_bin      1 1070.7
## - prothrombin    1 1074.0
## - albumin        1 1075.3
## - age_years      1 1077.1
## - log_bilirubin  1 1100.0
```

Table 1: Stepwise Cox Proportional Hazard Regression Model – coefficients

term	estimate	std.error	statistic	p.value	
edema_bin	0.476	0.231	2.058	0.040	*
albumin	-0.763	0.255	-2.992	0.003	**
copper	0.002	0.001	1.995	0.046	*
sgot	0.003	0.002	1.644	0.100	
prothrombin	0.289	0.103	2.806	0.005	**
stage2	1.346	1.061	1.269	0.204	
stage3	1.479	1.034	1.430	0.153	
stage4	1.763	1.033	1.707	0.088	
age_years	0.030	0.009	3.231	0.001	**
log_bilirubin	0.715	0.121	5.908	0.000	***

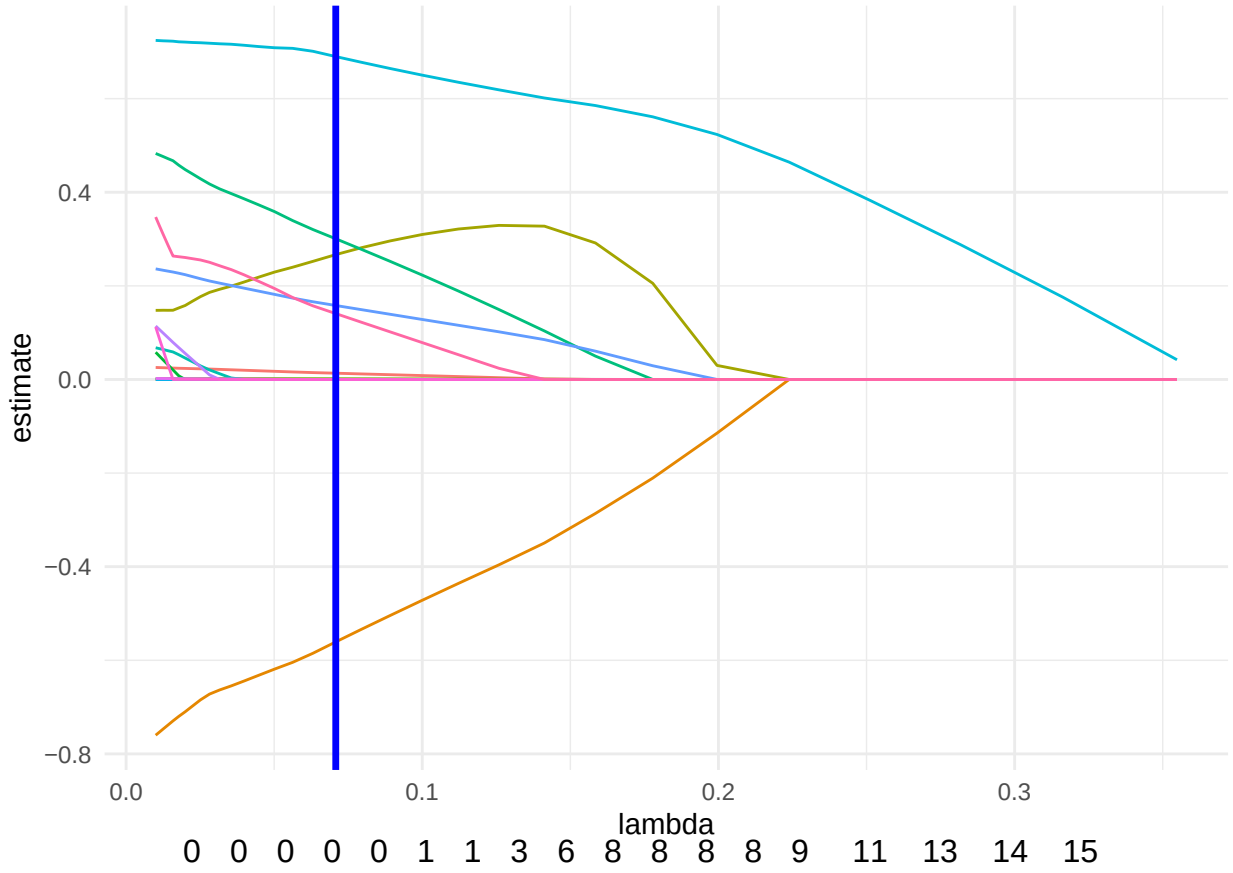
Table 2: Stepwise Cox Proportional Hazard Regression Model – Global Tests Results

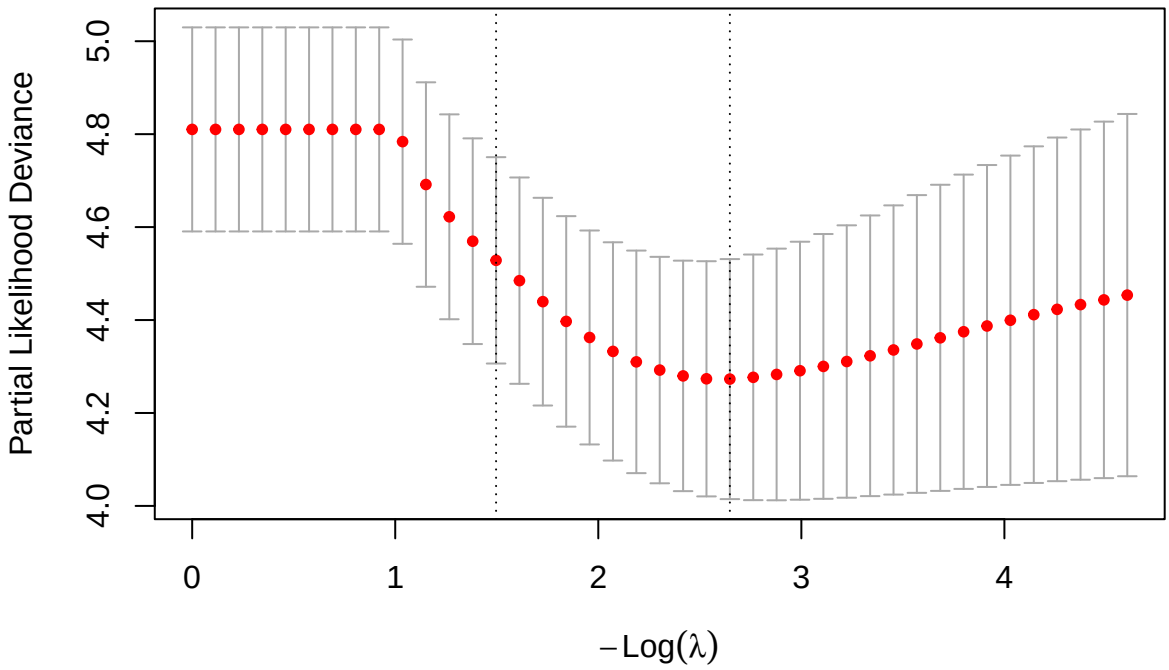
n	nevent	statistic.wald	p.value.wald	logLik	AIC	BIC
306	123	195.21	0	-524.311	1068.622	1096.744

A Stepwise Cox Proportional Hazard regression was performed to identify the most parsimonious model that estimates subjects' hazard of Cirrhosis. The model selection was based on Akaike's Information Criterion (AIC). The final model achieved the lowest AIC of 1068.622, including the variables of Edema Presence, Albumin(mg/dl), Urine copper(ug/day), SGOT (U/ml), Prothrombin time (s), histologic stage of disease, age (years) and log-transformed serum Bilirubin (mg/dl) as predictors. Several variables showed significant associations ($p < 0.05$) with the hazard function, while disease stages (all p-values> 0.05) and SGOT (p-value > 0.01) are weaker contributors to the model. The overall model shows significant discriminatory ability over subjects' hazards with a concordance of 0.858. Global tests show the model has a good fit (Wald test, p <2e-16).

The final model from stepwise selection further supported the results from the above data Cox ph model that drug is an insignificant predictor. Urine copper and SGOT, in addition, were kept in the stepwise Cox model, but were removed in the final proposed model.

Lasso Cox





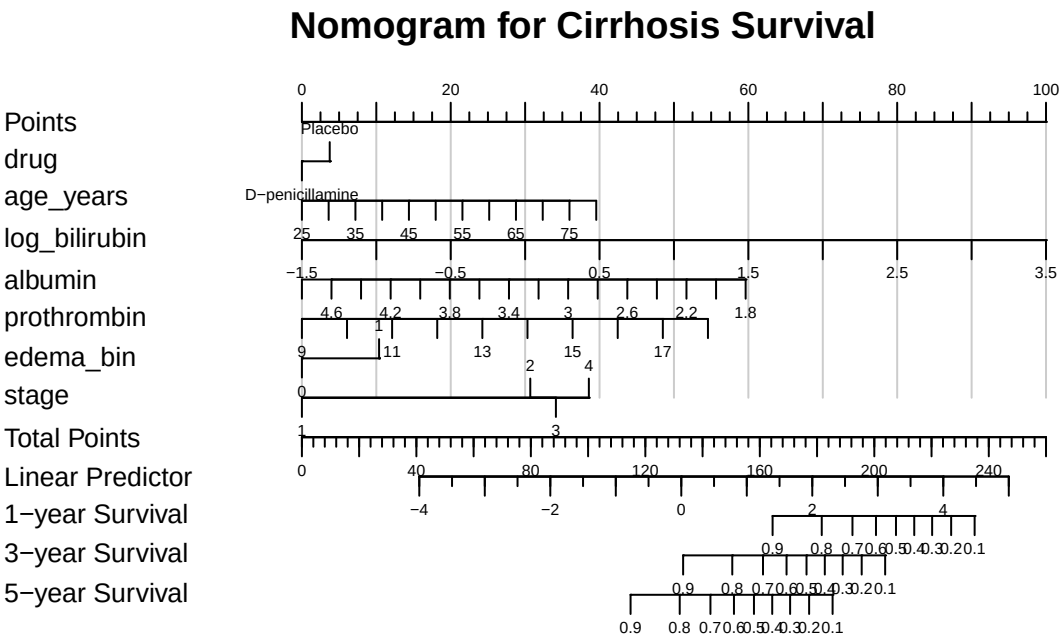
term	estimate_opt	estimate_1se
drugPlacebo	-	-

term	estimate_opt	estimate_1se
sexM	-	-
ascitesY	0.267	-
hepatomegalyY	-	-
spidersY	-	-
edema_bin	0.301	-
albumin	-0.561	-
copper	0.002	-
alk_phos	-	-
sgot	-	-
platelets	-	-
prothrombin	0.158	-
stage2	-	-
stage3	-	-
stage4	0.141	-
age_years	0.013	-
log_bilirubin	0.69	0.464

Another Lasso Cox Proportional Hazard regression was performed to identify the most parsimonious model that estimates subjects' hazard of Cirrhosis. The model selection was based on the Penalty Parameter, lambda. Cross-validation further selects the lambda that generalizes the best, and determines variables to develop the final model. The final model achieved the penalty parameter of 0.071, including the variables of Ascites Presence, Edema Presence, Albumin(mg/dl), Urine copper(ug/day), Prothrombin time (s), histologic stage 4 of disease, age (years) and log-transformed serum Bilirubin (mg/dl) as predictors. Another simpler model was provided with the log-transformed serum Bilirubin (mg/dl) as the only predictor, where allows the largest penalty parameter at which the MSE is within one standard error of the smallest MSE, which is 0.224. The model with the smallest penalty parameter was considered due to its better performance in estimation.

The final model from Lasso selection further supported the results from the above data Cox Proportional Hazard model that drug is an insignificant predictor. Ascites Presence and Urine copper, in addition, was kept in the Lasso Cox model, but was removed in the final proposed model.

Nomgram



1−, 3−, and 5−year survival predictions

The figure shows the nomogram generated to estimate patients' 1-, 3-, and 5-year survival based on the final Cox PH survival prediction model. The nomogram produces similar results when treatment types are not strong predictors in the model. It translates the regression coefficients into a point-based scoring system, allowing clinicians to estimate 1-, 3-, and 5-year survival probabilities for patients with cirrhosis. Each predictor in the model (drug assignment, age, log-bilirubin, albumin, prothrombin time, presence of edema, and disease stage) has a scale that assigns a value to a specific hazard point, as indicated by the scale at the top of the figure. By summing a patient's points for each variable, therapists can obtain a total score and use it to estimate 1-, 3-, and 5-year survival probabilities, with higher total point values indicating lower predicted survival. Nomograms provide an intuitive and feasible method for applying the survival prediction model and enhance clinical decision-making by offering quick, individualized risk estimates.