P8108 Group 2 Survival Analysis Project

Yiming Zhao (yz3955) Wenshan Qu (wq2160) Tucker Morgan (tlm2152)
Junzhe Shao (js5959) Benjamin Goebel (bpg2118)

2022-10-19

```
library(survival)
library(tidyverse)
library(tidymodels)
library(glmnet)
knitr::opts_chunk$set(message = FALSE, warning = FALSE)
```

Train Test Split

```
set.seed(2022)

rotterdam_split <- initial_split(rotterdam, prop = 0.8, strata = death)
rotterdam_training <- training(rotterdam_split)
rotterdam_test <- testing(rotterdam_split)
```

Perform 10-fold Cross-Validation

The output contains 1 row for each fold/repeat. So, 10 folds * 5 repeats = 50 rows. The split_analysis column is a list column containing a data frame for each row with 9 folds combined, and the split_assessment column is a list column containing a data frame for each row with 1 fold.

```
set.seed(2022)

rotterdam_folds <- vfold_cv(rotterdam_training, v = 10, repeats = 5,
                            strata = death)

rotterdam_folds <- rotterdam_folds %>%
  mutate(split_analysis = map(splits, analysis),
         split_assessment = map(splits, assessment))
```

Introduction

Methods

The dataset of interest for this analysis comes from the Rotterdam tumor bank, including data from 2982 breast cancer patients. Follow up time for patients varied from just 1 month to as long as 231 months. Several prognostic variables are recorded including year of surgery, age at surgery, menopausal status (pre-

or post-), tumor size (mm), differentiation grade, number of positive lymph nodes, progesterone receptors (fmol/l), estrogen receptors (fmol/l), and indicators for hormonal treatment and chemotherapy treatment. The outcome considered in this analysis was patient death.

(Placeholder for Cross-validation)

As part of this analysis, we consider the Cox Proportional Hazard (Cox PH) model, which allows us to model the hazard ratio based on covariates to understand their impact on the survival function. The Cox PH typically takes the form:

$$h(t|Z = z) = h_0(t)e^{\beta' z}.$$

In this application, we use the elastic net penalty, a mixture of the ℓ_1 and ℓ_2 norm regularization penalties. In the Cox PH framework, this penalty term takes the form of:

$$\lambda\left(\alpha \sum |\beta_i| + \frac{1}{2}(1-\alpha)\sum \beta_i^2\right)$$

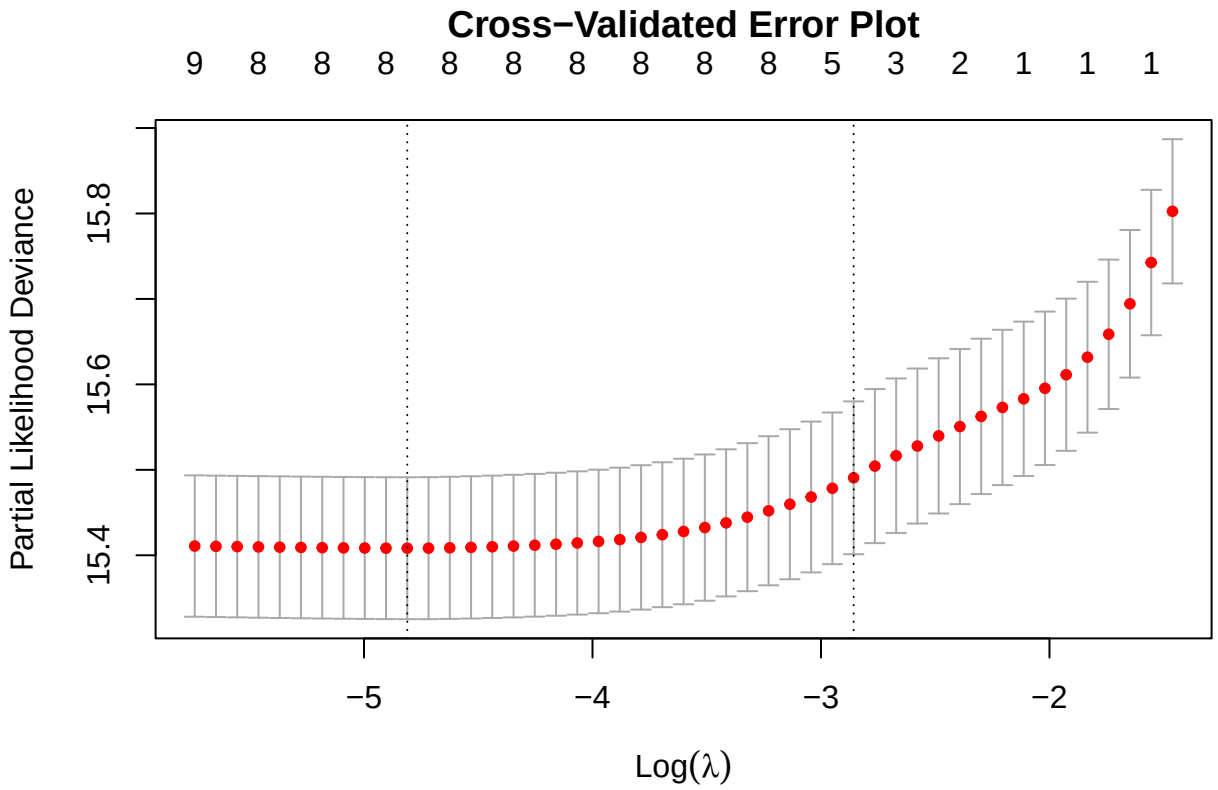
where λ represents our penalty coefficient and α is the mixing parameter for the two regularization methods. This penalty helps to avoid over-fitting of our data. The algorithm used here in `glmnet` uses the Breslow approximation to handle ties. For more details on the derivation of this term and the algorithm used to fit the penalized Cox PH model, see Simon et al. (2011).

Exploratory Data Analysis

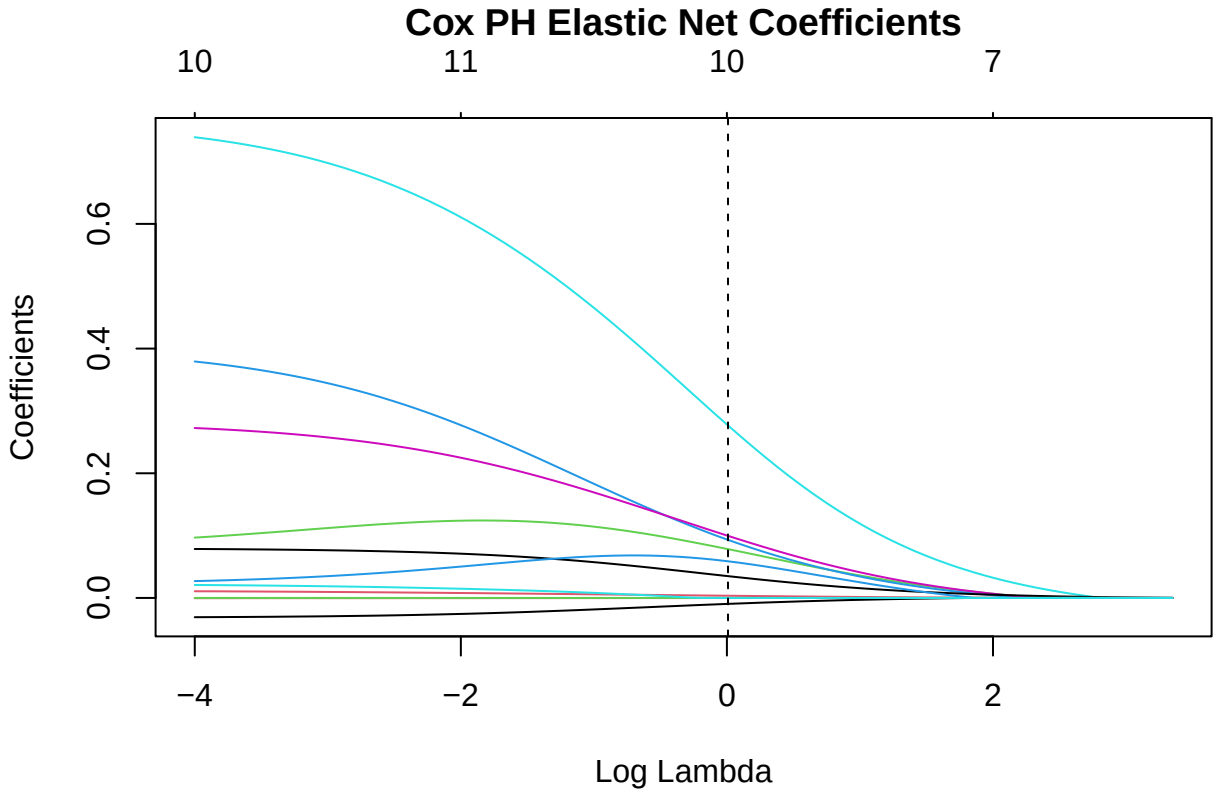
Cross-Validation

Cox/Cox with elastic net

```
set.seed(2022)
# removing relapse data, since death is the primary outcome
rotterdam_trn_d <-
  rotterdam_training %>%
  select(-rtime, -recur, -pid)

cox_trn_x <- model.matrix(Surv(dtime, death) ~ ., rotterdam_trn_d)[,-1]
cox_trn_y <- Surv(rotterdam_trn_d$dtime, rotterdam_trn_d$death)

cv_coxfit <- cv.glmnet(cox_trn_x, cox_trn_y, family = "cox", type.measure = "deviance")

par(mar = c(4,4,5,1))
plot(cv_coxfit, main = "Cross-Validated Error Plot")
```




```
coxfit <- glmnet(cox_trn_x, cox_trn_y, family = "cox", alpha = cv_coxfit$lambda.min)

par(mar = c(4,4,5,1))
plot(coxfit, xvar = "lambda",
     main = "Cox PH Elastic Net Coefficients")
abline(v = cv_coxfit$lambda.min, lty = 2)
```




```
coxfit_df <-
  data.frame(
    "coef" = as.vector(coef(coxfit, s = cv_coxfit$lambda.min)),
    "exp_coef" = as.vector(coef(coxfit, s = cv_coxfit$lambda.min)) %>% exp()
)

rownames(coxfit_df) <- labels(coef(coxfit, s = cv_coxfit$lambda.min))[[1]]

coxfit_df %>% round(digits = 4) %>%
  knitr::kable(caption = "Cox Proportion Hazard Elastic Net Coefficients")
```

Table 1: Cox Proportion Hazard Elastic Net Coefficients

	coef	exp_coef
year	-0.0309	0.9696
age	0.0108	1.0108
meno	0.0968	1.1016
size20-50	0.3793	1.4612
size>50	0.7391	2.0940
grade	0.2726	1.3134
nodes	0.0786	1.0817
pgr	-0.0004	0.9996
er	0.0000	1.0000
hormon	0.0270	1.0274

	coef	exp_coef
chemo	0.0208	1.0210

Note we are unable to provide standard errors for these estimates. Based on the results above, we see hazard ratios of:

- 0.9696 for year, a hazard reduction of 3.04% for a one-year increase in year of surgery;
- 1.0108 for age, a hazard increase of 1.08% for a one-year increase in patient age at surgery;
- 1.1016 for menopausal status, a hazard increase of 10.16% for a postmenopausal patient compared to a pre-menopausal patient, holding all else equal;
- 1.4612 for tumor size of 20-50 mm, a hazard increase of 46.12% for a tumor of size 20-50 mm compared to a tumor less than 20 mm;
- 2.094 for tumor size of greater than 50 mm, a hazard increase of 46.12% for a tumor of size greater than 50 units compared to a tumor less than 20 mm;
- 1.3134 for the differentiation grade, a hazard increase of 31.34% for a one-unit increase in differentiation grade;
- 1.0817 for the number of positive lymph nodes, a 8.17% hazard increase for each additional positive lymph node;
- 0.9996 for the progesterone receptors (fmol/l), a hazard reduction of 0.04% for a one-unit increase in progesterone receptors;
- 1 for estrogen receptors (fmol/l), a hazard change of 0% for each one-unit increase in estrogen receptors;
- 1.0274 for hormonal treatment, a hazard increase of 2.74% for patients who received hormonal therapy compared to patients who did not;
- 1.021 for chemotherapy treatment, a hazard increase of 2.1% for patients who received chemotherapy compared to patients who did not.

In our model, we find increased hazard treatment effects for both hormonal treatment and chemotherapy. While we might expect these treatments to reduce hazard, it is possible these results are confounded by treatment assignment. In other words, case severity might impact both treatment assignment and treatment response.

Random survival forest

Conformalized survival analysis

Supplemental analyses

Results

Discussion

How our results compare with past research

Conclusion

References

—Note this reference is in MLA format—

Simon, Noah et al. “Regularization Paths for Cox’s Proportional Hazards Model via Coordinate Descent.” Journal of statistical software vol. 39,5 (2011): 1-13. doi:10.18637/jss.v039.i05