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)
library(ranger)
library(survminer)
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:

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

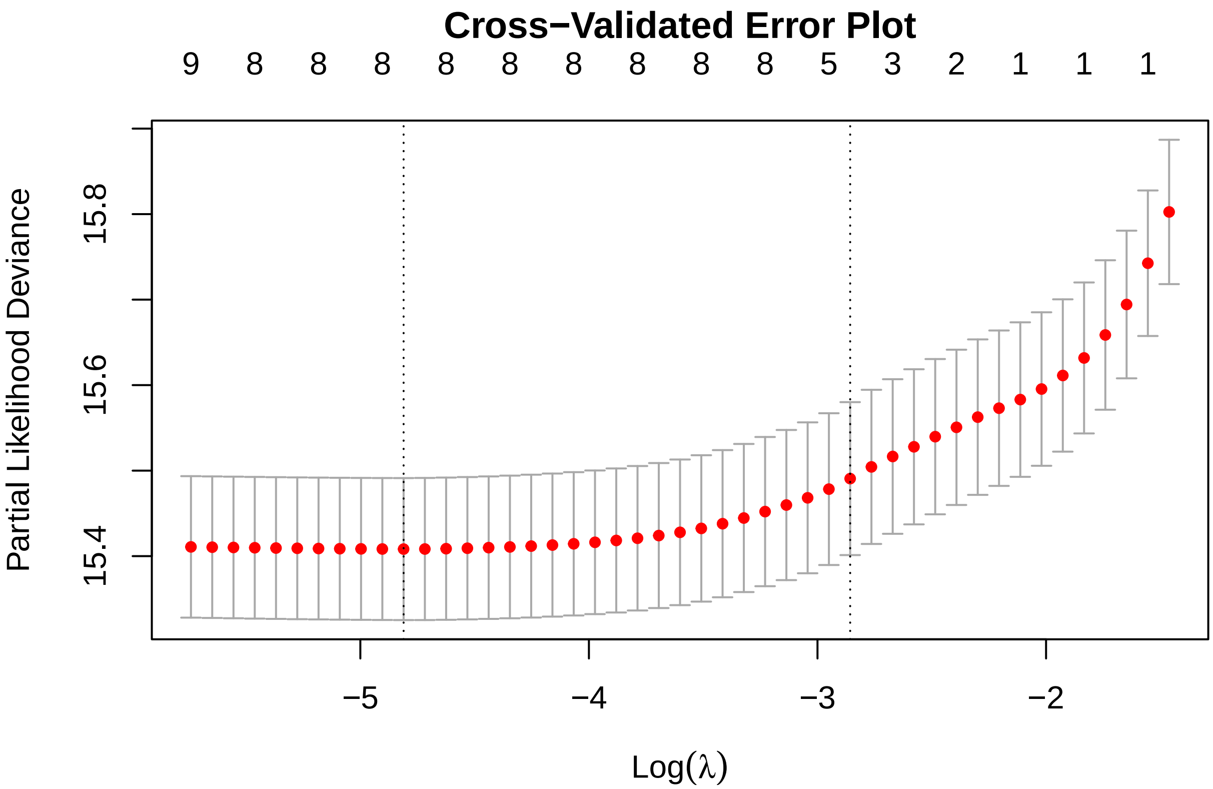
where λ represents our penalty coefficient and α is the mixing parameter for the two regularization methods. This penalty helps to avoid over-fitting of our data. The algorithm used here in `glmnet` uses the Breslow approximation to handle ties. For more details on the derivation of this term and the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

Exploratory Data Analysis

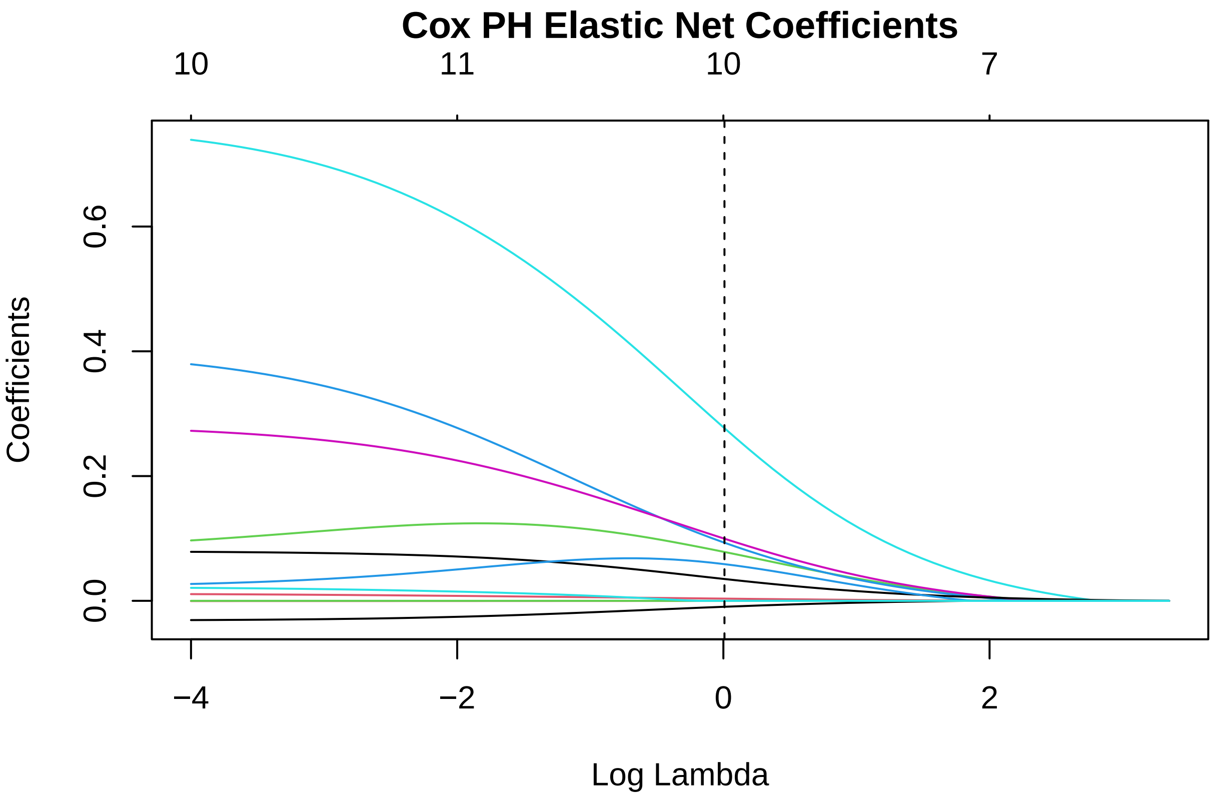
Cross-Validation

Cox/Cox with elastic net

```r
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
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.

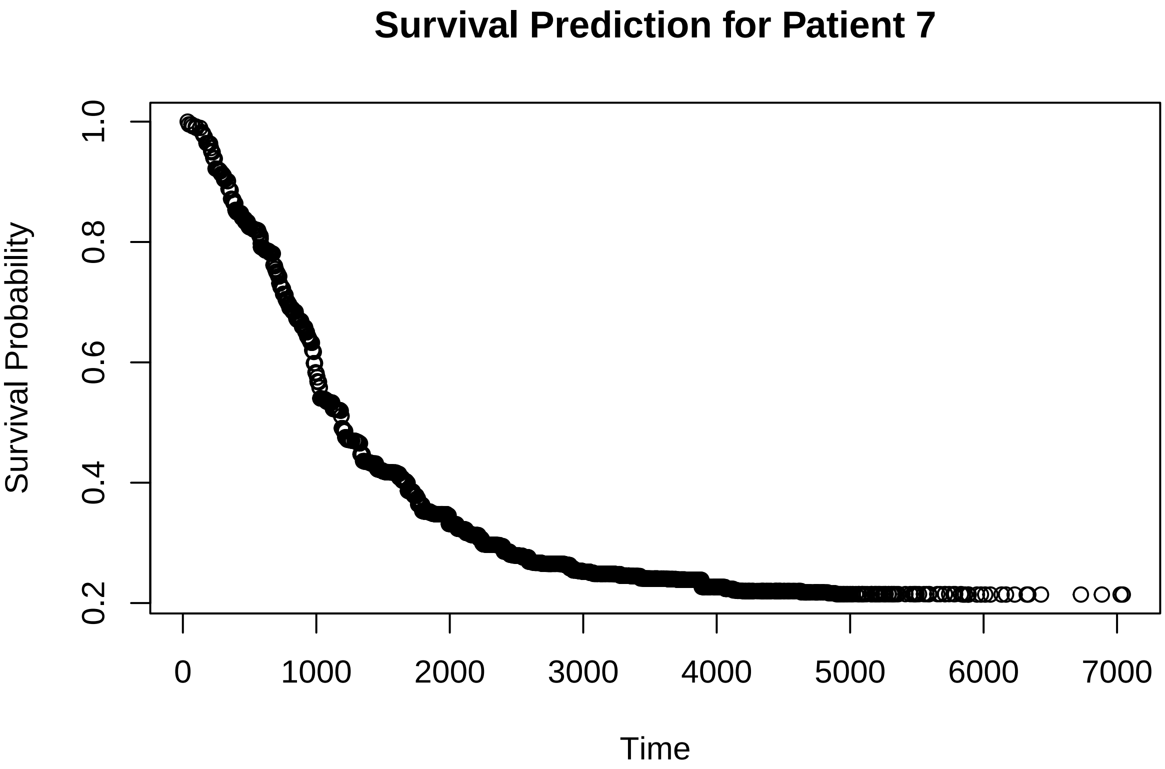
Random survival forest

The survival tree and the corresponding random survival forest (RSF) are highly favorable non-parametric methods when studying survival data. Generally, for a single survival tree, it will assign subjects to groups based on certain splitting rules regarding their covariates, and the subjects in each group will share a similar survival behavior.

```r
## Random Survival Forest
rsf <- ranger(Surv(time = dtime, event = death) ~ .,
              data = rotterdam_trn_d,
              num.trees = 300,
              min.node.size = 15)

## Remove variables not for prediction, and the outcome
rotterdam_test_d <-
  rotterdam_test %>%
  select(-rtime, -recur, -pid, -death)

## Make prediction on all the test data points
pred_rsf <- predict(rsf, rotterdam_test_d, type = "response")
# Look at individual 7
pred_ref_7 <- data.frame(
  time = pred_rsf$unique.death.times,
  survival = pred_rsf$survival[7,])
head(pred_ref_7) %>% knitr::kable(align = "c")
```

time	survival
36	1.0000000
45	0.9954515
64	0.9954515
74	0.9923594
97	0.9920546
101	0.9896016

```
plot(pred_ref_7$time, pred_ref_7$survival,
     xlab = "Time", ylab = "Survival Probability",
     main = "Survival Prediction for Patient 7")
```



```
# Find estimated median survival time for individual 7
head(pred_ref_7[pred_ref_7$survival <= 0.5,]) %>% knitr::kable(align = "c") #1163
```

	time	survival
346	1191	0.4906214
347	1192	0.4906214
348	1193	0.4906214
349	1198	0.4906214
350	1200	0.4875678
351	1204	0.4856753

```
# See the truth of individual 7
rotterdam_test[7,] %>% knitr::kable(align = "c")
```

	pid	year	age	meno	size	grade	nodes	pgr	er	hormon	chemo	rtime	recur	dtime	death
2463	58	1992	69	1	20-50	2	8	5	6	1	0	1869	0	1869	0

With `ranger` package, we trained the random survival forest with training dataset used for survival prediction. As a non-parametric method, there is no parameters in RSF that could be interpreted. The ultimate goal of RSF is to predict the survival probability function of a given data point based on its covariate vector. Compared to semi-parametric Cox-PH model which forces the outcome and the covariates to have a special connection, the RSF makes prediction based on the survival time of training data points that shares similar propensity with the given input data point.

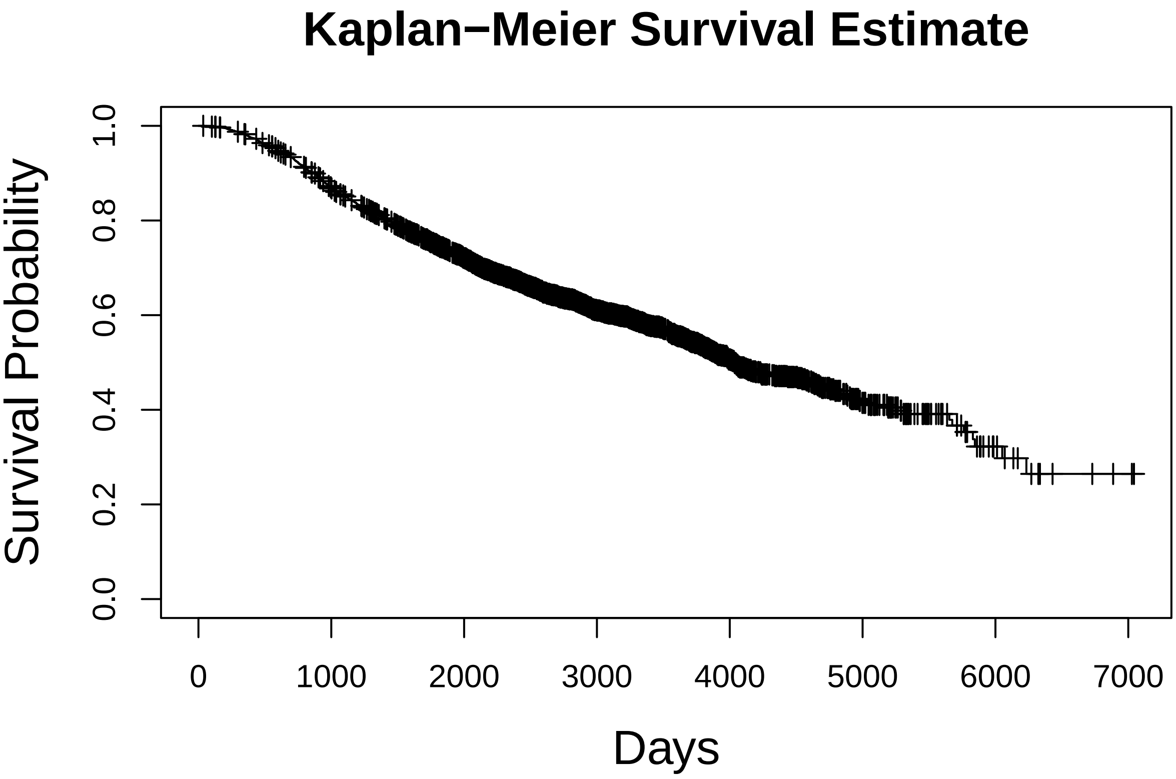
Since the "truth" of test data point (a single survival time) and the prediction we made here (a survival probability function) are not comparable, here we show the prediction result of the 7th test data point (pid = 58). The survival curve has been shown above, and the median survival time is 1163 days.

Conformalized survival analysis

Supplemental analyses

Kaplan-Meier Survival Estimate

```
KM = survfit(Surv(dtime, death) ~ 1, data = rotterdam)
plot(KM, conf.int = FALSE, mark.time = TRUE,
     xlab = "Days", ylab = "Survival Probability",
     main = "Kaplan-Meier Survival Estimate", cex.lab = 1.5, cex.main = 1.5)
```



Log-rank Test

The null hypothesis of our log-rank test is: $H_0 : S_1(t) = S_0(t)$, where $S_1(t)$ is the survival function of hormon treatment group, $S_0(t)$ is the survival function of control group.

```
logrank <- survdiff(Surv(dtime, death) ~ hormon, data = rotterdam)
logrank
```

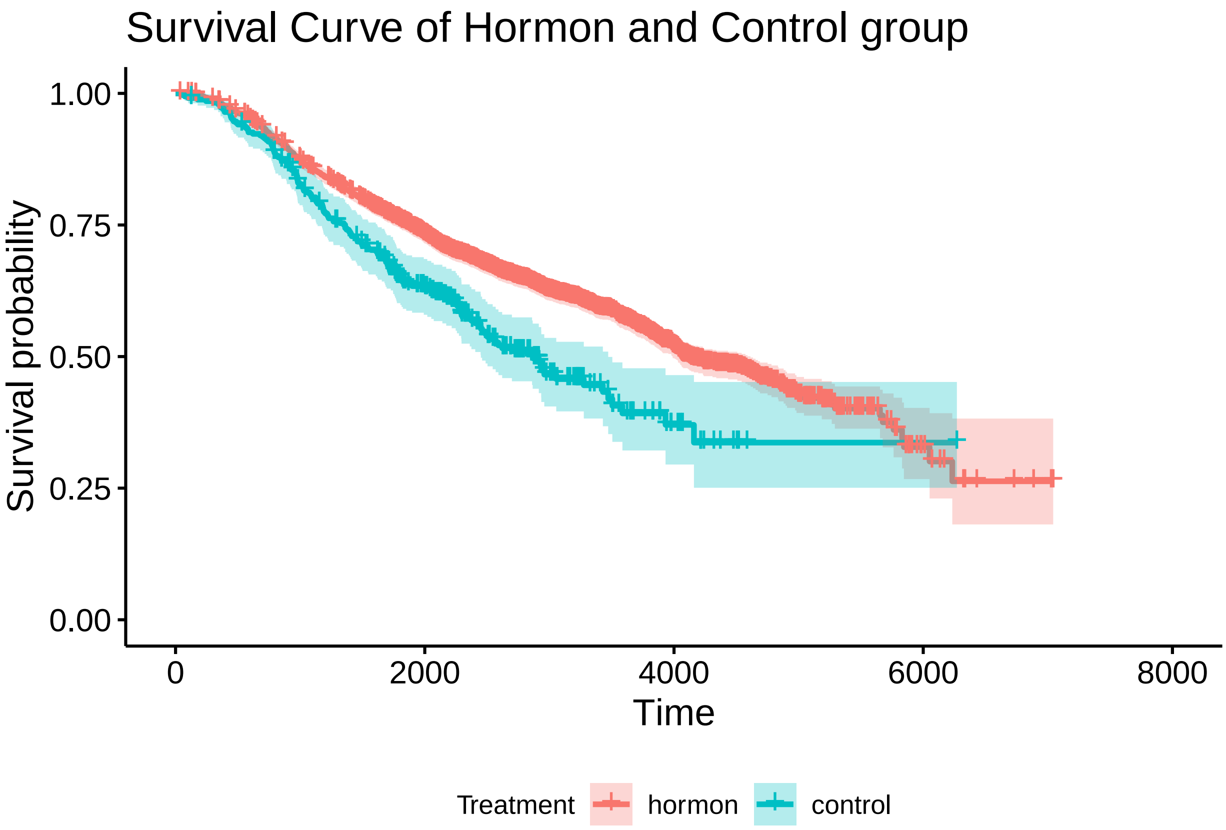
```
## Call:
## survdiff(formula = Surv(dtime, death) ~ hormon, data = rotterdam)
##
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## hormon=0 2643     1113     1162      2.04      23.7
## hormon=1  339      159      110     21.43      23.7
##
##  Chisq= 23.7  on 1 degrees of freedom, p= 1e-06
```

```
logrank$pvalue
```

```
## [1] 1.133649e-06
```

The test statistic is 23.7, and the corresponding p-value is $1.133^{-6} \ll 0.05$, thus we reject the null and conclude that we are 95% confident that $S_1(t) \neq S_0(t$. And since the test statistic is positive, we can conclude that the hormon treatment is significantly effective to breast cancer.

```
ggsurvplot(survfit(Surv(dtime,death) ~ hormon, data = rotterdam),
           conf.int = TRUE,
           legend = c("bottom"),
           legend.title = c("Treatment"),
           legend.labs = c("hormon", "control")) +
  ggtitle("Survival Curve of Hormon and Control group")
```



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