Advanced Survival Analysis

A Comprehensive Study of Cure Models

Andrew Kamya

December, 2024

Introduction

Survival analysis is a cornerstone of statistical methods used to analyze time-to-event data, particularly in medical research. Traditional survival models assume that all individuals in the population are susceptible to the event of interest (e.g., death, relapse). However, in many real-world scenarios, a subset of individuals may be “cured” or no longer at risk of the event. Cure models extend traditional survival analysis by incorporating the possibility of a cured fraction, making them particularly useful in cancer research and other fields where long-term survival is observed.

This report explores advanced topics in survival analysis, focusing on **cure models**. We analyze three datasets using parametric and semiparametric cure models, evaluate their performance, and interpret the results. The analysis is divided into three parts, each addressing a specific dataset and research question.

Methodology

We employ both parametric and semiparametric cure models to analyze survival data. Parametric models assume specific distributions for the survival times (e.g., Weibull, Gamma, Log-normal), while semiparametric models, such as the proportional hazards (PH) mixture cure model, do not require such assumptions. The models are fitted using R packages such as `flexsurv`, `smcure`, and `survival`.

Key steps in the analysis include:

1. Data Preparation: Loading and exploring the datasets.
2. Model Fitting: Fitting parametric and semiparametric cure models.
3. Model Selection: Using AIC to select the best-fitting model.
4. Interpretation: Analyzing the coefficients and cure fractions.
5. Visualization: Plotting survival curves and predicted cure rates.

Part I: Breast Cancer Survival Data Analysis

Data Description

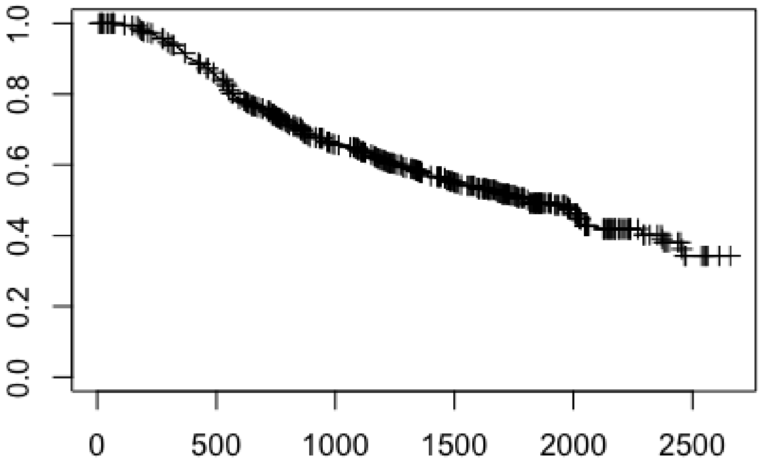
The bc dataset, available in the flexsurv package, contains survival times of 686 patients with primary node-positive breast cancer. The variables include: - censrec: Event indicator (1 = dead, 0 = censored). - rectime: Time of death or censoring in days. - group: Prognostic group (Good, Medium, Poor).

Analysis

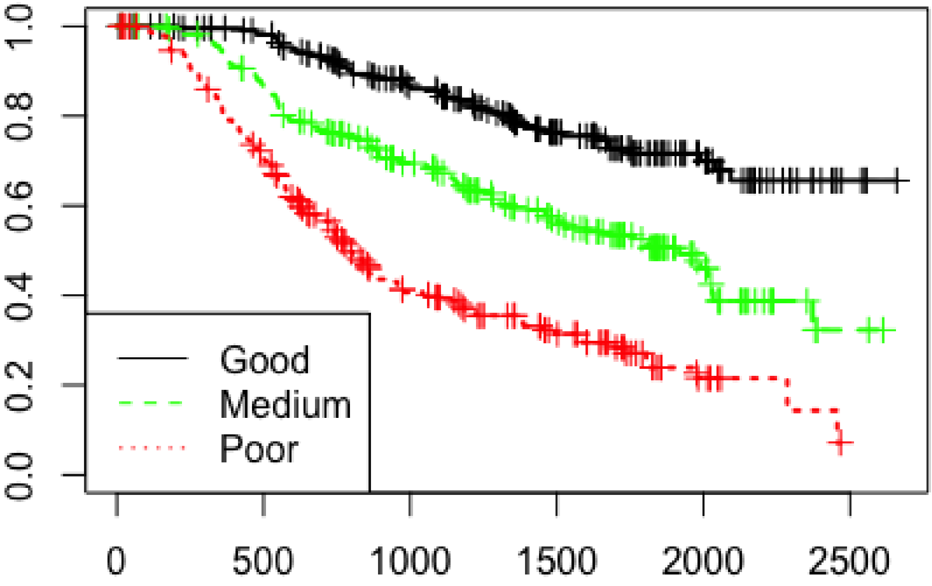
Step 1: Exploratory Analysis

We begin by visualizing the survival curves for the entire dataset and by prognostic group.

```
library("flexsurv")

## Loading required package: survival

library("flexsurvcure")

data("bc")

s1 <- survfit(Surv(rectime, censrec) ~ 1, data = bc)
plot(s1, mark.time = TRUE, conf.int = FALSE)
```



```r
s2 <- survfit(Surv(rectime, censrec) ~ group, data = bc)
plot(s2, col = c('black', 'green', 'red'), lty = 1:3, mark.time = TRUE, lwd =
2)
legend('bottomleft', c("Good", "Medium", "Poor"), lty = 1:3, col = c('black',
'green', 'red'))
```



Step 2: Fitting Parametric Cure Models

We fit parametric cure models with Weibull, Gamma, Exponential, and Log-normal distributions for the latency part and logistic regression for the incidence part.

```r
para_bc_gamma <- flexsurvcure(Surv(rectime, censrec) ~ group, data = bc,
                              anc = list(rate = ~group), dist = "gamma", link
= "logistic", mixture = TRUE)

para_bc_weibell <- flexsurvcure(Surv(rectime, censrec) ~ group, data = bc,
                                anc = list(scale = ~group), dist = "weibull",
link = "logistic", mixture = TRUE)

para_bc_lnorm <- flexsurvcure(Surv(rectime, censrec) ~ group, data = bc,
                              anc = list(meanlog = ~group), dist = "lnorm",
link = "logistic", mixture = TRUE)
```

```
para_bc_exp <- flexsurvcure(Surv(rectime, censrec) ~ group, data = bc,
                            anc = list(rate = ~group), dist = "exp", link =
"logistic", mixture = TRUE)

AIC(para_bc_gamma, para_bc_weibell, para_bc_lnorm, para_bc_exp)
```

```
##                  df      AIC
## para_bc_gamma    7 5138.150
## para_bc_weibell  7 5153.300
## para_bc_lnorm    7 5119.421
## para_bc_exp      6 5202.351
```

Step 3: Interpretation

The Log-normal model has the lowest AIC and is selected as the best-fitting model. The results indicate: - Incidence: Prognostic group significantly impacts the probability of being cured. - Since the "flexsurvcure" uses logistic regression to model the probability of being cured, odds ratio of exp(-0.6138)=0.541 to be cured for median prognostic group compared to the good prognostic group. - OR of exp(-1.585)= 0.205 to be cured for poor prognostic group compared to the good prognostic group.

- Latency: Prognostic group does not significantly affect the survival time of uncured patients (confidence intervals include 0).
- The mean log survival time in median prognostic group is 0.524 days. Shorter than the mean log survival time in good prognostic group.
- The mean log survival time in poor prognostic group is 1.008 days. Shorter than the mean log survival time in good prognostic group.

Part II: Clinical Trial Data Analysis

Data Description

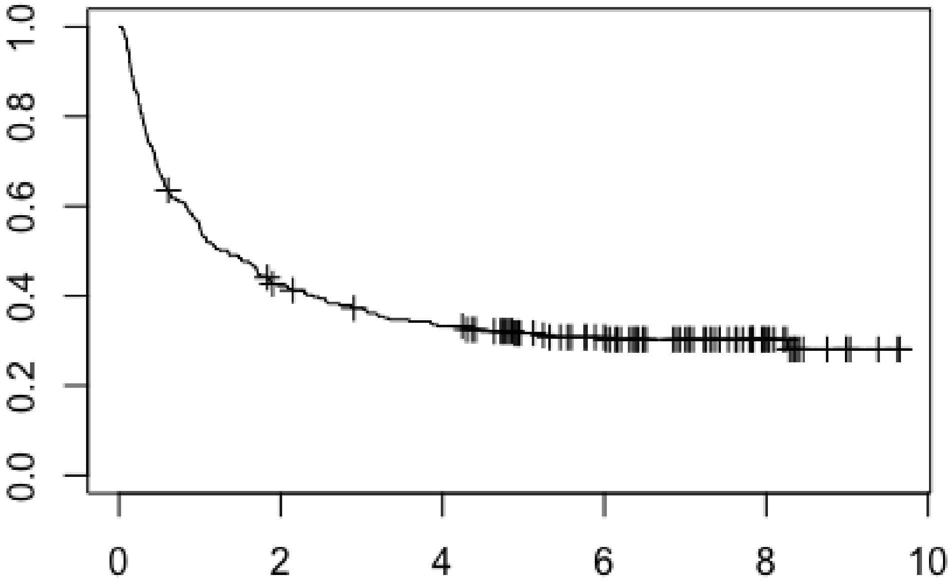
The `e1684` dataset, available in the `smcure` package, contains data from a clinical trial evaluating high-dose interferon alpha-2b as postoperative adjuvant therapy. The variables include: - `TRT`: Treatment group (0 = control, 1 = treatment). - `SEX`: Gender (0 = male, 1 = female). - `AGE`: Age (centered to the mean). - `FAILTIME`: Relapse-free survival time in years. - `FAILCENS`: Event indicator (1 = relapse, 0 = censored).

Analysis

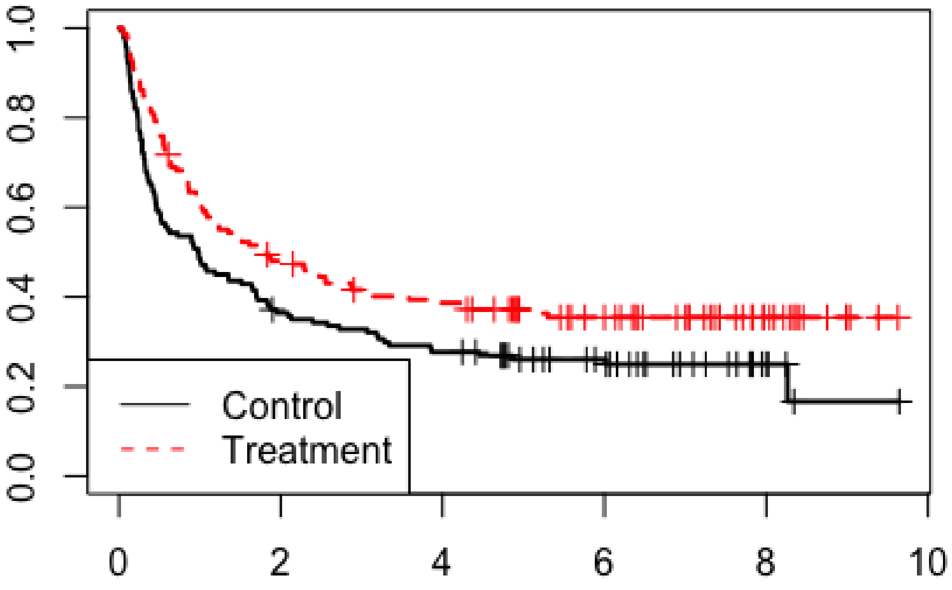
Step 1: Kaplan-Meier Estimators

We estimate and plot the overall survival curves and survival curves by treatment and gender.

```
library("survival")
library("smcure")

data(e1684)

E2_s <- survfit(Surv(FAILTIME, FAILCENS) ~ 1, data = e1684)
plot(E2_s, mark.time = TRUE, conf.int = FALSE)
```



```
E2_treat <- survfit(Surv(FAILTIME, FAILCENS) ~ TRT, data = e1684)
plot(E2_treat, col = c('black', 'red'), lty = 1:2, mark.time = TRUE, lwd = 2)
legend('bottomleft', c("Control", "Treatment"), lty = 1:2, col = c('black',
'red'))
```



-

Patients in the treatment group have higher cure rate than those in control group. - Survival curves show that the difference in the cure rates between male and female patients is not significant. —

Step 2: Semiparametric PH Mixture Cure Model

We fit a semiparametric proportional hazards mixture cure model using treatment, gender, and age as covariates.

```
sm.ph <- smcure(Surv(FAILTIME, FAILCENS) ~ TRT + SEX + AGE, cureform = ~TRT +
SEX + AGE, data = e1684, model = "ph")

## Program is running..be patient... done.
## Call:
## smcure(formula = Surv(FAILTIME, FAILCENS) ~ TRT + SEX + AGE,
##     cureform = ~TRT + SEX + AGE, data = e1684, model = "ph")
##
## Cure probability model:
##                Estimate  Std.Error    Z value     Pr(>|Z|)
## (Intercept)  1.36493298 0.31476648  4.3363352 0.0000144878
## TRT         -0.58847727 0.35975826 -1.6357575 0.1018903470
## SEX         -0.08696490 0.31855018 -0.2730022 0.7848515344
## AGE          0.02033857 0.01734803  1.1723846 0.2410426596
##
```

```
##
## Failure time distribution model:
##         Estimate   Std.Error    Z value  Pr(>|Z|)
## TRT -0.153595097 0.194260445 -0.7906658 0.4291390
## SEX  0.099458470 0.183179429  0.5429565 0.5871597
## AGE -0.007664013 0.006424724 -1.1928936 0.2329110
```

```
printsmcure(sm.ph)
```

```
## Call:
## smcure(formula = Surv(FAILTIME, FAILCENS) ~ TRT + SEX + AGE,
##     cureform = ~TRT + SEX + AGE, data = e1684, model = "ph")
##
## Cure probability model:
##                Estimate  Std.Error    Z value     Pr(>|Z|)
## (Intercept)  1.36493298 0.31476648  4.3363352 0.0000144878
## TRT         -0.58847727 0.35975826 -1.6357575 0.1018903470
## SEX         -0.08696490 0.31855018 -0.2730022 0.7848515344
## AGE          0.02033857 0.01734803  1.1723846 0.2410426596
##
##
## Failure time distribution model:
##         Estimate   Std.Error    Z value  Pr(>|Z|)
## TRT -0.153595097 0.194260445 -0.7906658 0.4291390
## SEX  0.099458470 0.183179429  0.5429565 0.5871597
## AGE -0.007664013 0.006424724 -1.1928936 0.2329110
```

Step 3: Bootstrap Variance Estimation

We estimate the variance of the coefficients using bootstrap resampling.

```
sm.ph100 <- smcure(Surv(FAILTIME, FAILCENS) ~ TRT + SEX + AGE, cureform =
~TRT + SEX + AGE, data = e1684, model = "ph", nboot = 100)
```

```
## Program is running..be patient... done.
## Call:
## smcure(formula = Surv(FAILTIME, FAILCENS) ~ TRT + SEX + AGE,
##     cureform = ~TRT + SEX + AGE, data = e1684, model = "ph",
##     nboot = 100)
##
## Cure probability model:
##                Estimate  Std.Error    Z value     Pr(>|Z|)
## (Intercept)  1.36493298 0.30682172  4.4486192 8.642409e-06
## TRT         -0.58847727 0.32104307 -1.8330166 6.680010e-02
## SEX         -0.08696490 0.29739967 -0.2924176 7.699674e-01
## AGE          0.02033857 0.01351021  1.5054223 1.322155e-01
##
##
## Failure time distribution model:
##         Estimate   Std.Error    Z value  Pr(>|Z|)
## TRT -0.153595097 0.144886148 -1.0601089 0.2890951
## SEX  0.099458470 0.182933277  0.5436871 0.5866568
## AGE -0.007664013 0.006709278 -1.1423007 0.2533291
```

```
sm.ph200 <- smcure(Surv(FAILTIME, FAILCENS) ~ TRT + SEX + AGE, cureform =
~TRT + SEX + AGE, data = e1684, model = "ph", nboot = 200)
```

```
## Program is running..be patient... done.
## Call:
## smcure(formula = Surv(FAILTIME, FAILCENS) ~ TRT + SEX + AGE,
##     cureform = ~TRT + SEX + AGE, data = e1684, model = "ph",
##     nboot = 200)
##
## Cure probability model:
##                Estimate  Std.Error    Z value     Pr(>|Z|)
## (Intercept)  1.36493298 0.32048372  4.2589776 0.0000205364
## TRT         -0.58847727 0.35751675 -1.6460132 0.0997610478
## SEX         -0.08696490 0.34992088 -0.2485273 0.8037264402
## AGE          0.02033857 0.01648615  1.2336756 0.2173237936
##
##
## Failure time distribution model:
##         Estimate   Std.Error    Z value  Pr(>|Z|)
## TRT -0.153595097 0.182898636 -0.8397826 0.4010303
## SEX  0.099458470 0.188911564  0.5264816 0.5985536
## AGE -0.007664013 0.007270785 -1.0540832 0.2918448
```

```
sm.ph500 <- smcure(Surv(FAILTIME, FAILCENS) ~ TRT + SEX + AGE, cureform =
~TRT + SEX + AGE, data = e1684, model = "ph", nboot = 500)
```

```
## Program is running..be patient... done.
## Call:
## smcure(formula = Surv(FAILTIME, FAILCENS) ~ TRT + SEX + AGE,
##     cureform = ~TRT + SEX + AGE, data = e1684, model = "ph",
##     nboot = 500)
##
## Cure probability model:
##                Estimate  Std.Error    Z value     Pr(>|Z|)
## (Intercept)  1.36493298 0.32275353  4.2290258 2.347055e-05
## TRT         -0.58847727 0.34458001 -1.7078102 8.767156e-02
## SEX         -0.08696490 0.34785039 -0.2500066 8.025822e-01
## AGE          0.02033857 0.01485799  1.3688637 1.710419e-01
##
##
## Failure time distribution model:
##         Estimate   Std.Error    Z value  Pr(>|Z|)
## TRT -0.153595097 0.180938087 -0.8488821 0.3959469
## SEX  0.099458470 0.177631351  0.5599151 0.5755374
## AGE -0.007664013 0.006578407 -1.1650256 0.2440086
```

```
printsmcure(sm.ph100); printsmcure(sm.ph200); printsmcure(sm.ph500)
```

```
## Call:
## smcure(formula = Surv(FAILTIME, FAILCENS) ~ TRT + SEX + AGE,
##     cureform = ~TRT + SEX + AGE, data = e1684, model = "ph",
##     nboot = 100)
##
## Cure probability model:
##                Estimate  Std.Error    Z value     Pr(>|Z|)
## (Intercept)  1.36493298 0.30682172  4.4486192 8.642409e-06
## TRT         -0.58847727 0.32104307 -1.8330166 6.680010e-02
## SEX         -0.08696490 0.29739967 -0.2924176 7.699674e-01
## AGE          0.02033857 0.01351021  1.5054223 1.322155e-01
##
##
## Failure time distribution model:
##         Estimate   Std.Error    Z value  Pr(>|Z|)
## TRT -0.153595097 0.144886148 -1.0601089 0.2890951
## SEX  0.099458470 0.182933277  0.5436871 0.5866568
## AGE -0.007664013 0.006709278 -1.1423007 0.2533291
```

```
## Call:
## smcure(formula = Surv(FAILTIME, FAILCENS) ~ TRT + SEX + AGE,
##     cureform = ~TRT + SEX + AGE, data = e1684, model = "ph",
##     nboot = 200)
##
## Cure probability model:
##                Estimate  Std.Error    Z value     Pr(>|Z|)
## (Intercept)  1.36493298 0.32048372  4.2589776 0.0000205364
## TRT         -0.58847727 0.35751675 -1.6460132 0.0997610478
## SEX         -0.08696490 0.34992088 -0.2485273 0.8037264402
## AGE          0.02033857 0.01648615  1.2336756 0.2173237936
##
##
## Failure time distribution model:
##         Estimate   Std.Error    Z value  Pr(>|Z|)
## TRT -0.153595097 0.182898636 -0.8397826 0.4010303
## SEX  0.099458470 0.188911564  0.5264816 0.5985536
## AGE -0.007664013 0.007270785 -1.0540832 0.2918448
```

```
## Call:
## smcure(formula = Surv(FAILTIME, FAILCENS) ~ TRT + SEX + AGE,
##     cureform = ~TRT + SEX + AGE, data = e1684, model = "ph",
##     nboot = 500)
##
## Cure probability model:
##                Estimate  Std.Error    Z value     Pr(>|Z|)
## (Intercept)  1.36493298 0.32275353  4.2290258 2.347055e-05
## TRT         -0.58847727 0.34458001 -1.7078102 8.767156e-02
## SEX         -0.08696490 0.34785039 -0.2500066 8.025822e-01
## AGE          0.02033857 0.01485799  1.3688637 1.710419e-01
##
```

```
##
## Failure time distribution model:
##         Estimate   Std.Error    Z value  Pr(>|Z|)
## TRT -0.153595097 0.180938087 -0.8488821 0.3959469
## SEX  0.099458470 0.177631351  0.5599151 0.5755374
## AGE -0.007664013 0.006578407 -1.1650256 0.2440086
```

Part III: Bone Marrow Transplant Data Analysis

Data Description

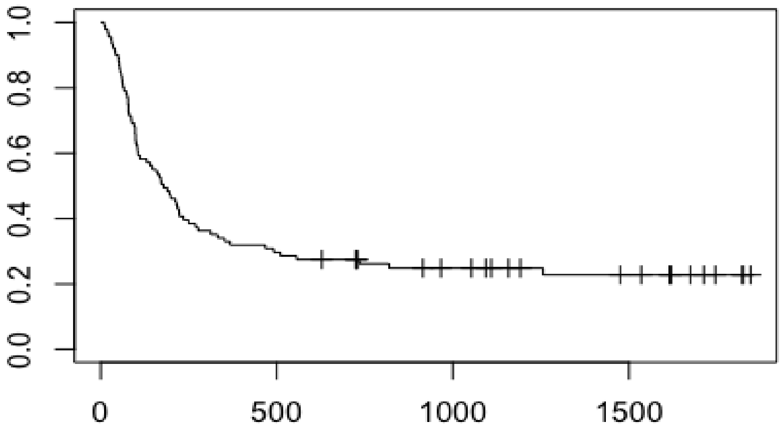
The bmt dataset, available in the smcure package, contains data from a bone marrow transplant study. The variables include: - TRT: Treatment group (0 = allogeneic, 1 = autologous). - Time: Time to death. - Status: Event indicator (1 = death, 0 = censored).

Analysis

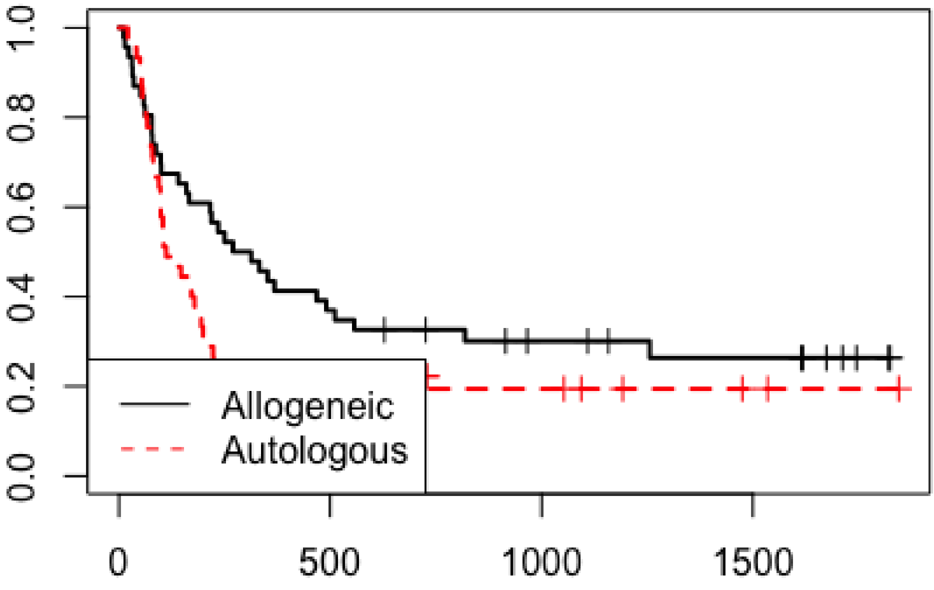
Step 1: Kaplan-Meier Estimators

We estimate and plot the overall survival curves and survival curves by treatment group.

```
data(bmt, package = "smcure")

E3_s1 <- survfit(Surv(Time, Status) ~ 1, data = bmt)
plot(E3_s1, mark.time = TRUE, conf.int = FALSE)
```



```
E3_s2 <- survfit(Surv(Time, Status) ~ TRT, data = bmt)
plot(E3_s2, col = c('black', 'red'), lty = 1:2, mark.time = TRUE, lwd = 2)
legend('bottomleft', c("Allogeneic", "Autologous"), lty = 1:2, col =
c('black', 'red'))
```



Step 2: Semiparametric PH and AFT Mixture Cure Models

We fit semiparametric PH and AFT mixture cure models.

```
sm.ph <- smcure(Surv(Time, Status) ~ TRT, cureform = ~TRT, data = bmt, model
= "ph", Var = TRUE)

## Program is running..be patient... done.
## Call:
## smcure(formula = Surv(Time, Status) ~ TRT, cureform = ~TRT, data = bmt,
##     model = "ph", Var = TRUE)
##
## Cure probability model:
##              Estimate Std.Error   Z value     Pr(>|Z|)
## (Intercept) 1.0565750 0.2758690 3.8299882 0.0001281494
## TRT         0.3579095 0.5598042 0.6393478 0.5225967316
##
##
## Failure time distribution model:
```

```
##      Estimate Std.Error  Z value  Pr(>|Z|)
## TRT 0.6363645 0.3472718 1.832468 0.0668817

# sm.aft <- smcure(Surv(Time, Status) ~ TRT, cureform = ~TRT, data = bmt,
model = "aft", Var = FALSE)
```

Conclusion

This report demonstrates the application of cure models in survival analysis using three distinct datasets. Key findings include: 1. **Breast Cancer Data**: The Log-normal cure model provided the best fit, with prognostic group significantly impacting the cure probability. 2. **Clinical Trial Data**: Treatment and gender did not significantly affect the cure probability or survival time. 3. **Bone Marrow Transplant Data**: The allogeneic treatment group showed a higher cure rate compared to the autologous group.

Cure models are powerful tools for analyzing survival data with a cured fraction, providing insights that traditional survival models cannot capture. Future work could explore more complex models and larger datasets to further validate these findings.