harinris_Homework6

Harin Rishabh

2024-05-04

Lab 11.8

11.8.1-Brain Cancer Data

```r
library(ISLR2)
```

```r
names(BrainCancer)
```

```
## [1] "sex"       "diagnosis" "loc"       "ki"        "gtv"       "stereo"
## [7] "status"    "time"
```

```r
attach(BrainCancer)
table(sex)
```

```
## sex
## Female    Male
##     45      43
```

```r
table(diagnosis)
```

```
## diagnosis
## Meningioma  LG glioma  HG glioma      Other
##         42          9         22         14
```

```r
table(status)
```

```
## status
##  0  1
## 53 35
```

```r
length(BrainCancer$time)
```

```
## [1] 88
```

```r
length(BrainCancer$status)
```

```
## [1] 88
```

```
class(BrainCancer$time)
```

```
## [1] "numeric"
```

```
class(BrainCancer$status)
```

```
## [1] "integer"
```

```
str(BrainCancer)
```

```
## 'data.frame':    88 obs. of  8 variables:
## $ sex      : Factor w/ 2 levels "Female","Male": 1 2 1 1 2 1 2 2 1 2 ...
## $ diagnosis: Factor w/ 4 levels "Meningioma","LG glioma",..: 1 3 1 2 3 1 1 2 1 3
## ...
## $ loc      : Factor w/ 2 levels "Infratentorial",..: 1 2 1 2 2 2 2 2 2 2 ...
## $ ki       : int  90 90 70 80 90 80 80 80 70 100 ...
## $ gtv      : num  6.11 19.35 7.95 7.61 5.06 ...
## $ stereo   : Factor w/ 2 levels "SRS","SRT": 1 2 1 2 2 1 2 2 2 2 ...
## $ status   : int  0 1 0 1 1 0 0 0 0 ...
## $ time     : num  57.64 8.98 26.46 47.8 6.3 ...
```

```
head(BrainCancer)
```

```
##       sex  diagnosis           loc ki   gtv stereo status  time
## 1 Female Meningioma Infratentorial 90  6.11    SRS      0 57.64
## 2   Male  HG glioma Supratentorial 90 19.35    SRT      1  8.98
## 3 Female Meningioma Infratentorial 70  7.95    SRS      0 26.46
## 4 Female  LG glioma Supratentorial 80  7.61    SRT      1 47.80
## 5   Male  HG glioma Supratentorial 90  5.06    SRT      1  6.30
## 6 Female Meningioma Supratentorial 80  4.82    SRS      0 52.75
```

```
library(survival)

surv_obj <- Surv(BrainCancer$time, BrainCancer$status)

fit.surv <- survfit(surv_obj ~ 1)

plot(fit.surv, xlab = "Months",
ylab = "Estimated Probability of Survival")
```
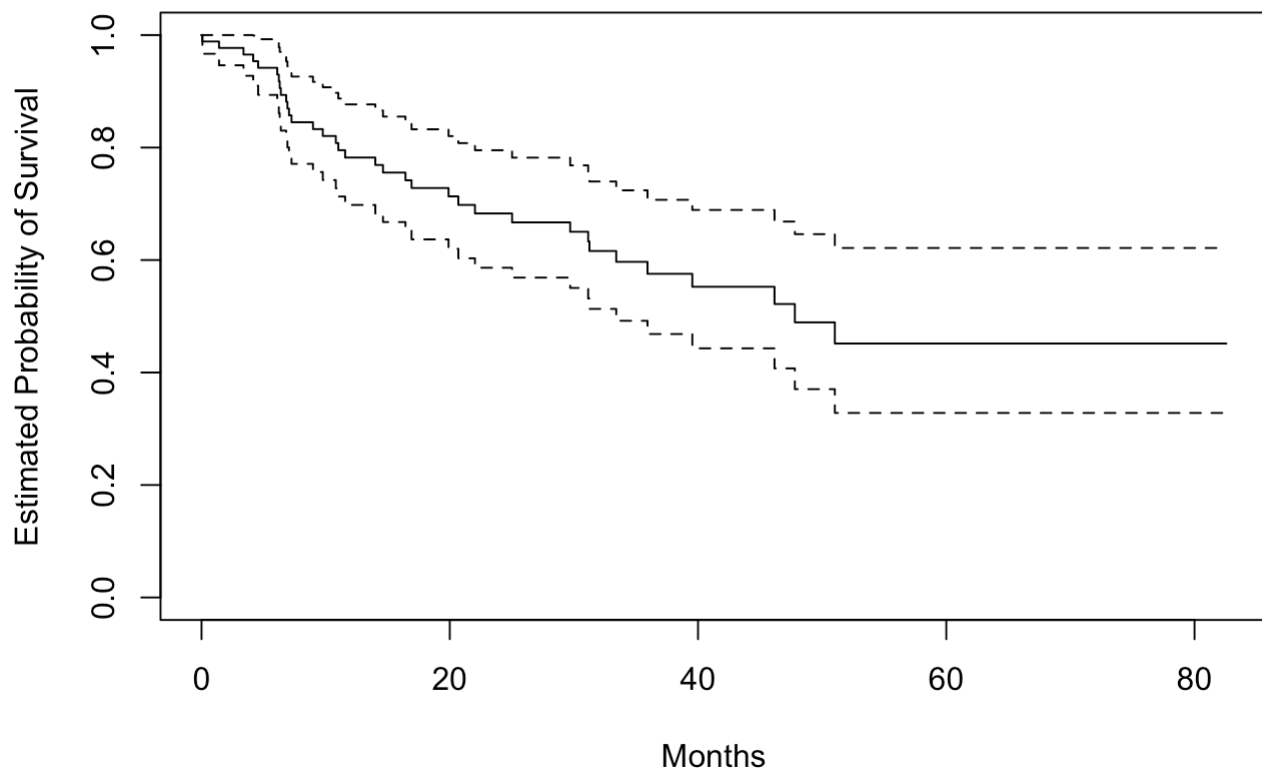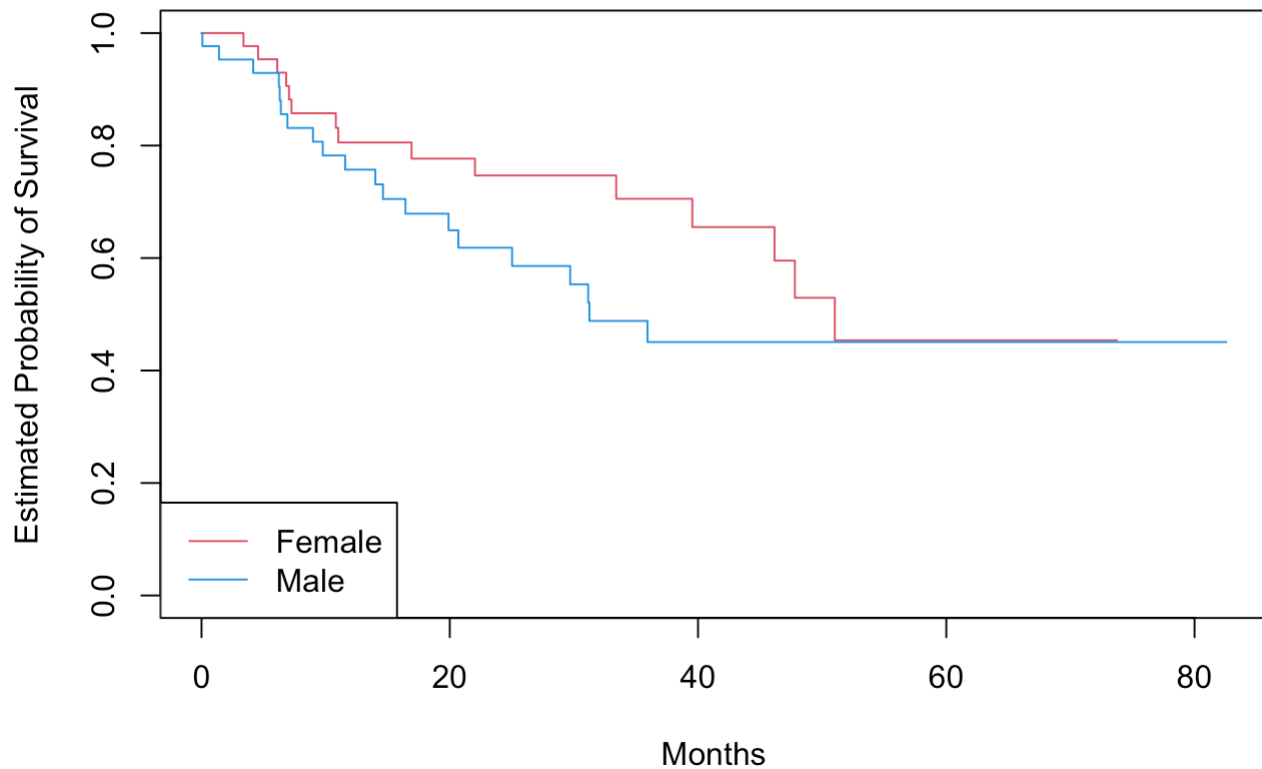
```r
BrainCancer$sex <- factor(BrainCancer$sex)

fit.sex <- survfit(Surv(time, status) ~ sex, data = BrainCancer)

plot(fit.sex, xlab = "Months", ylab = "Estimated Probability of Survival", col = c(2,
4))

legend("bottomleft", legend = levels(BrainCancer$sex), col = c(2, 4), lty = 1)
```

```
logrank.test <- survdiff(Surv(time, status) ~ sex, data = BrainCancer)
logrank.test
```

```
## Call:
## survdiff(formula = Surv(time, status) ~ sex, data = BrainCancer)
##
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## sex=Female 45       15     18.5     0.676      1.44
## sex=Male   43       20     16.5     0.761      1.44
##
##  Chisq= 1.4  on 1 degrees of freedom, p= 0.2
```

```
fit.cox <- coxph(Surv(time, status) ~ sex, data = BrainCancer)
summary(fit.cox)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ sex, data = BrainCancer)
##
##   n= 88, number of events= 35
##
##            coef exp(coef) se(coef)     z Pr(>|z|)
## sexMale 0.4077    1.5033   0.3420 1.192    0.233
##
##          exp(coef) exp(-coef) lower .95 upper .95
## sexMale     1.503     0.6652     0.769     2.939
##
## Concordance= 0.565  (se = 0.045 )
## Likelihood ratio test= 1.44  on 1 df,   p=0.2
## Wald test            = 1.42  on 1 df,   p=0.2
## Score (logrank) test = 1.44  on 1 df,   p=0.2
```

```
summary(fit.cox)$logtest[1]
```

```
##     test
## 1.438822
```

```
summary(fit.cox)$waldtest[1]
```

```
## test
## 1.42
```

```
summary(fit.cox)$sctest[1]
```

```
##     test
## 1.440495
```

```
logrank.test$chisq
```

```
## [1] 1.440495
```

```
fit.all <- coxph(Surv(time, status) ~ sex + diagnosis + loc + ki + gtv + stereo, data
= BrainCancer)

fit.all
```

```
## Call:
## coxph(formula = Surv(time, status) ~ sex + diagnosis + loc +
##     ki + gtv + stereo, data = BrainCancer)
##
##                         coef exp(coef) se(coef)      z        p
## sexMale              0.18375   1.20171  0.36036  0.510  0.61012
## diagnosisLG glioma   0.91502   2.49683  0.63816  1.434  0.15161
## diagnosisHG glioma   2.15457   8.62414  0.45052  4.782 1.73e-06
## diagnosisOther       0.88570   2.42467  0.65787  1.346  0.17821
## locSupratentorial    0.44119   1.55456  0.70367  0.627  0.53066
## ki                  -0.05496   0.94653  0.01831 -3.001  0.00269
## gtv                  0.03429   1.03489  0.02233  1.536  0.12466
## stereoSRT            0.17778   1.19456  0.60158  0.296  0.76760
##
## Likelihood ratio test=41.37  on 8 df, p=1.776e-06
## n= 87, number of events= 35
##    (1 observation deleted due to missingness)
```

```
modaldata <- data.frame(
diagnosis = levels(diagnosis),
sex = rep("Female", 4),
loc = rep("Supratentorial", 4),
ki = rep(mean(ki), 4),
gtv = rep(mean(gtv), 4),
stereo = rep("SRT", 4)
)
```

```
survplots <- survfit(fit.all, newdata = modaldata)
plot(survplots, xlab = "Months",
ylab = "Survival Probability", col = 2:5)
legend("bottomleft", levels(diagnosis), col = 2:5, lty = 1)
```
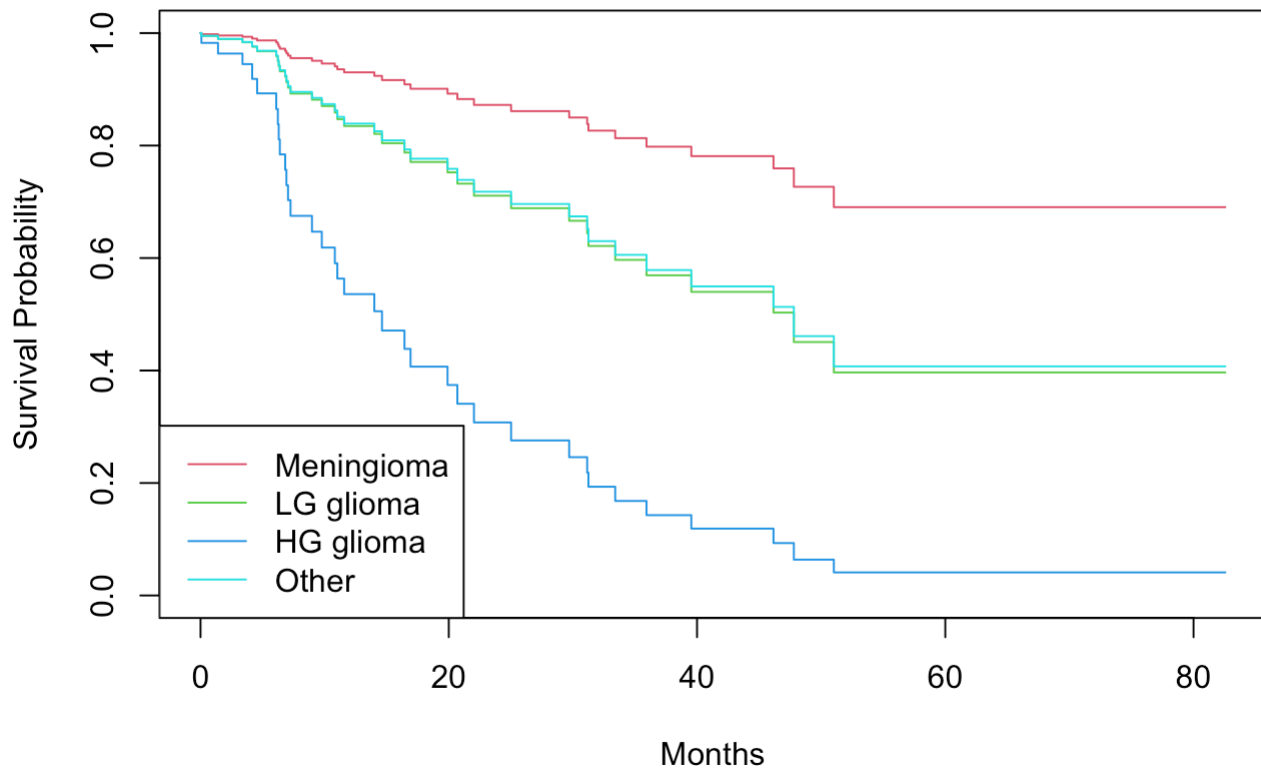
11.8.2 - Publication Data

```r
library(survival)
library(ISLR2)
fit.posres <- survfit(Surv(time, status) ~ posres, data = Publication)
```

```r
plot(fit.posres, xlab = "Months",
ylab = "Probability of Not Being Published", col = 3:4)
legend("topright", c("Negative Result", "Positive Result"),
col = 3:4, lty = 1)
```
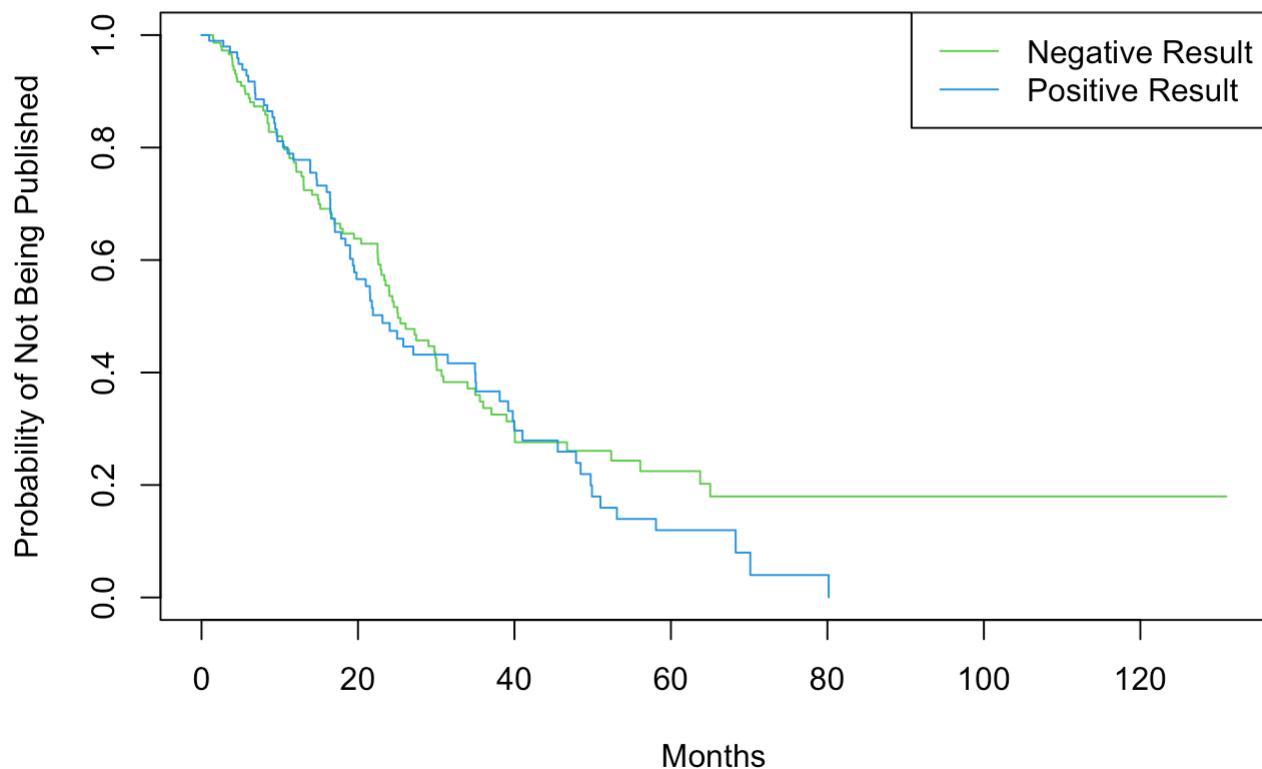
```
fit.pub <- coxph(Surv(time, status) ~ posres, data = Publication)
summary(fit.pub)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ posres, data = Publication)
##
##   n= 244, number of events= 156
##
##          coef exp(coef) se(coef)     z Pr(>|z|)
## posres 0.1481    1.1596   0.1616 0.916     0.36
##
##        exp(coef) exp(-coef) lower .95 upper .95
## posres      1.16     0.8624    0.8448     1.592
##
## Concordance= 0.503  (se = 0.022 )
## Likelihood ratio test= 0.83  on 1 df,   p=0.4
## Wald test            = 0.84  on 1 df,   p=0.4
## Score (logrank) test = 0.84  on 1 df,   p=0.4
```

```
logrank.test <- survdiff(Surv(time, status) ~ posres, data = Publication)
logrank.test
```

```
## Call:
## survdiff(formula = Surv(time, status) ~ posres, data = Publication)
##
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## posres=0 146       87     92.6     0.341     0.844
## posres=1  98       69     63.4     0.498     0.844
##
##  Chisq= 0.8  on 1 degrees of freedom, p= 0.4
```

```
fit.pub2 <- coxph(Surv(time, status) ~ . - mech, data = Publication)
summary(fit.pub2)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ . - mech, data = Publication)
##
##   n= 244, number of events= 156
##
##                 coef  exp(coef)  se(coef)       z Pr(>|z|)
## posres    5.708e-01  1.770e+00  1.760e-01   3.244  0.00118 **
## multi    -4.086e-02  9.600e-01  2.512e-01  -0.163  0.87079
## clinend   5.462e-01  1.727e+00  2.620e-01   2.085  0.03710 *
## sampsize  4.678e-06  1.000e+00  1.472e-05   0.318  0.75070
## budget    4.385e-03  1.004e+00  2.465e-03   1.779  0.07518 .
## impact    5.832e-02  1.060e+00  6.676e-03   8.735  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## posres       1.770     0.5651    1.2534     2.498
## multi        0.960     1.0417    0.5867     1.571
## clinend      1.727     0.5792    1.0332     2.885
## sampsize     1.000     1.0000    1.0000     1.000
## budget       1.004     0.9956    0.9996     1.009
## impact       1.060     0.9433    1.0463     1.074
##
## Concordance= 0.801  (se = 0.016 )
## Likelihood ratio test= 149.2  on 6 df,   p=<2e-16
## Wald test            = 159.3  on 6 df,   p=<2e-16
## Score (logrank) test = 233.7  on 6 df,   p=<2e-16
```

11.8.3 - Call Center Data

```
set.seed(4)

N <- 2000

Operators <- sample(5:15, N, replace = TRUE)

Center <- sample(c("A", "B", "C"), N, replace = TRUE)

Time <- sample(c("Morn.", "After.", "Even."), N, replace = TRUE)

X <- model.matrix(~ Operators + Center + Time)[, -1]
```

```
X[1:5, ]
```

```
##   Operators CenterB CenterC TimeEven. TimeMorn.
## 1        12       1       0         0         1
## 2        15       0       0         0         0
## 3         7       0       1         1         0
## 4         7       0       0         0         0
## 5        11       0       1         0         1
```

```
true.beta <- c(0.04, -0.3, 0, 0.2, -0.2)
h.fn <- function(x) return(0.00001 * x)
```

```
library(rms)
```

```
## Warning: package 'rms' was built under R version 4.3.2
```

```
## Loading required package: Hmisc
```

```
## Warning: package 'Hmisc' was built under R version 4.3.2
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
library(coxed)
```

```
## Loading required package: mgcv
```

```
## Loading required package: nlme
```

```
## This is mgcv 1.9-1. For overview type 'help("mgcv-package")'.
```

```
queuing <- sim.survdata(N = N, T = 1000, X = X, beta = true.beta, hazard.fun = h.fn)
```

```
## Warning in FUN(X[[i]], ...): 9 additional observations right-censored because the
user-supplied hazard function
##                                  is nonzero at the latest timepoint. To avoid the
se extra censored observations, increase T
```

```
names(queuing)
```

```
## [1] "data"              "xdata"             "baseline"          "xb"
## [5] "exp.xb"            "betas"             "ind.survive"       "marg.effect"
## [9] "marg.effect.data"
```

```
head(queuing$data)
```

```
##   Operators CenterB CenterC TimeEven. TimeMorn.   y failed
## 1        12       1       0         0         1 344   TRUE
## 2        15       0       0         0         0 241   TRUE
## 3         7       0       1         1         0 187   TRUE
## 4         7       0       0         0         0 279   TRUE
## 5        11       0       1         0         1 954   TRUE
## 6         7       1       0         0         1 455   TRUE
```

```
mean(queuing$data$failed)
```

```
## [1] 0.89
```

```
par(mfrow = c(1, 2))

fit.Center <- survfit(Surv(y, failed) ~ Center, data = queuing$data)

plot(fit.Center, xlab = "Seconds",
     ylab = "Probability of Still Being on Hold",
     col = c(2, 4, 5))  # Use different colors for each 'Center'

legend("topright",
       c("Call Center A", "Call Center B", "Call Center C"),  # Legend labels
       col = c(2, 4, 5),  # Colors corresponding to each 'Center'
       lty = 1)  # Line type (solid)

par(mfrow = c(1, 1))
```
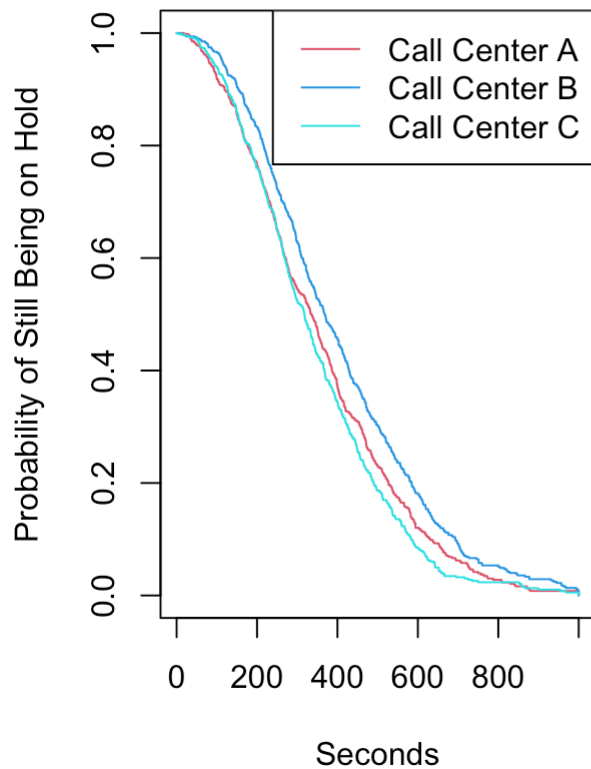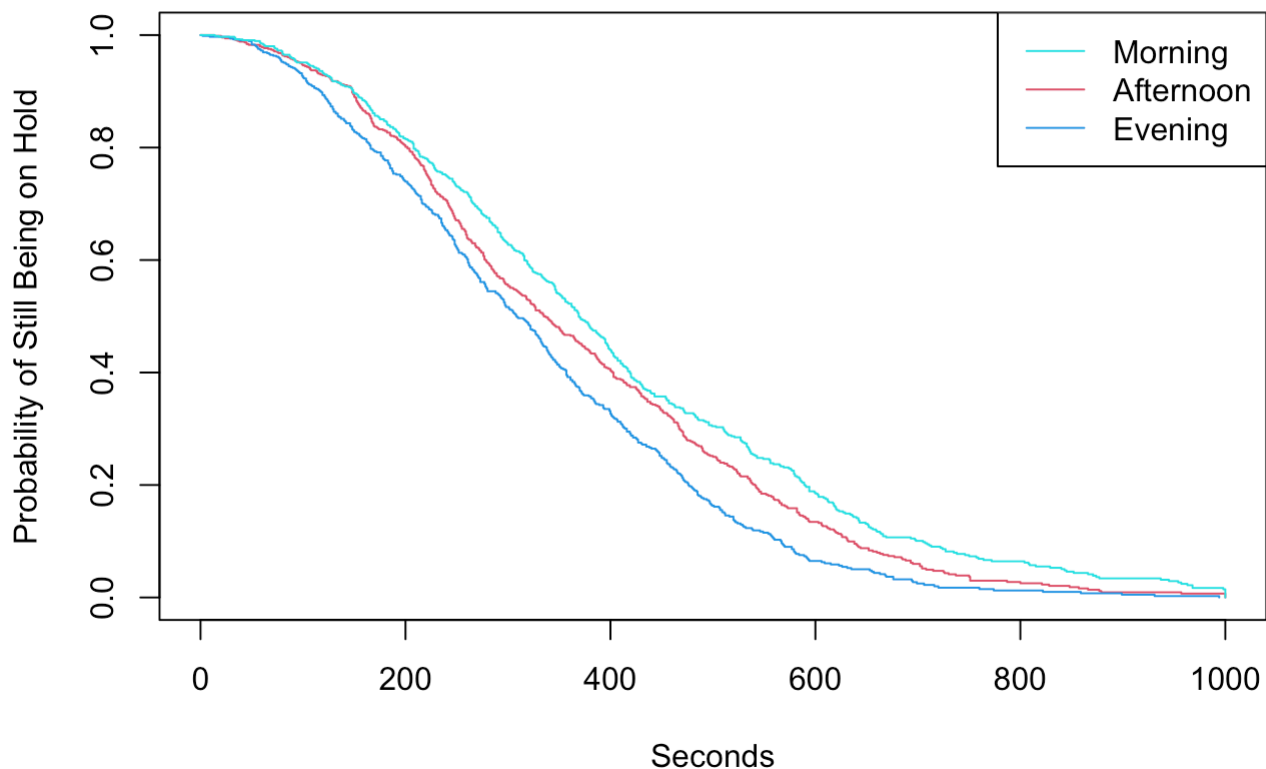
```
fit.Time <- survfit(Surv(y, failed) ~ Time, data = queuing$data)

plot(fit.Time, xlab = "Seconds",
     ylab = "Probability of Still Being on Hold",
     col = c(2, 4, 5))  # Use different colors for each 'Time'

legend("topright",
       c("Morning", "Afternoon", "Evening"),  # Legend labels
       col = c(5, 2, 4),  # Colors corresponding to each 'Time'
       lty = 1)  # Line type (solid)
```

```
survdiff(Surv(y, failed) ~ Center, data = queuing$data)
```

```
## Call:
## survdiff(formula = Surv(y, failed) ~ Center, data = queuing$data)
##
##            N Observed Expected (O-E)^2/E (O-E)^2/V
## Center=A 683      603      579     0.971      1.45
## Center=B 667      600      701    14.641     24.64
## Center=C 650      577      499    12.062     17.05
##
##  Chisq= 28.3  on 2 degrees of freedom, p= 7e-07
```

```
survdiff(Surv(y, failed) ~ Time, data = queuing$data)
```

```
## Call:
## survdiff(formula = Surv(y, failed) ~ Time, data = queuing$data)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## Time=After. 688      616      619    0.0135     0.021
## Time=Even.  653      582      468   27.6353    38.353
## Time=Morn.  659      582      693   17.7381    29.893
##
##  Chisq= 46.8  on 2 degrees of freedom, p= 7e-11
```

```
fit.queuing <- coxph(Surv(y, failed) ~ ., data = queuing$data)

fit.queuing
```

```
## Call:
## coxph(formula = Surv(y, failed) ~ ., data = queuing$data)
##
##                 coef exp(coef) se(coef)      z        p
## Operators   0.04174   1.04263  0.00759  5.500  3.8e-08
## CenterB    -0.21879   0.80349  0.05793 -3.777 0.000159
## CenterC     0.07930   1.08253  0.05850  1.356 0.175256
## TimeEven.   0.20904   1.23249  0.05820  3.592 0.000328
## TimeMorn.  -0.17352   0.84070  0.05811 -2.986 0.002828
##
## Likelihood ratio test=102.8  on 5 df, p=< 2.2e-16
## n= 2000, number of events= 1780
```

Applied Problem 10

10. a.

```
library(ISLR2)
x <- Surv(BrainCancer$time, BrainCancer$status)
plot(survfit(x ~ 1),
  xlab = "Months",
  ylab = "Estimated Probability of Survival",
  col = "blue",
  conf.int = 0.67
)
```
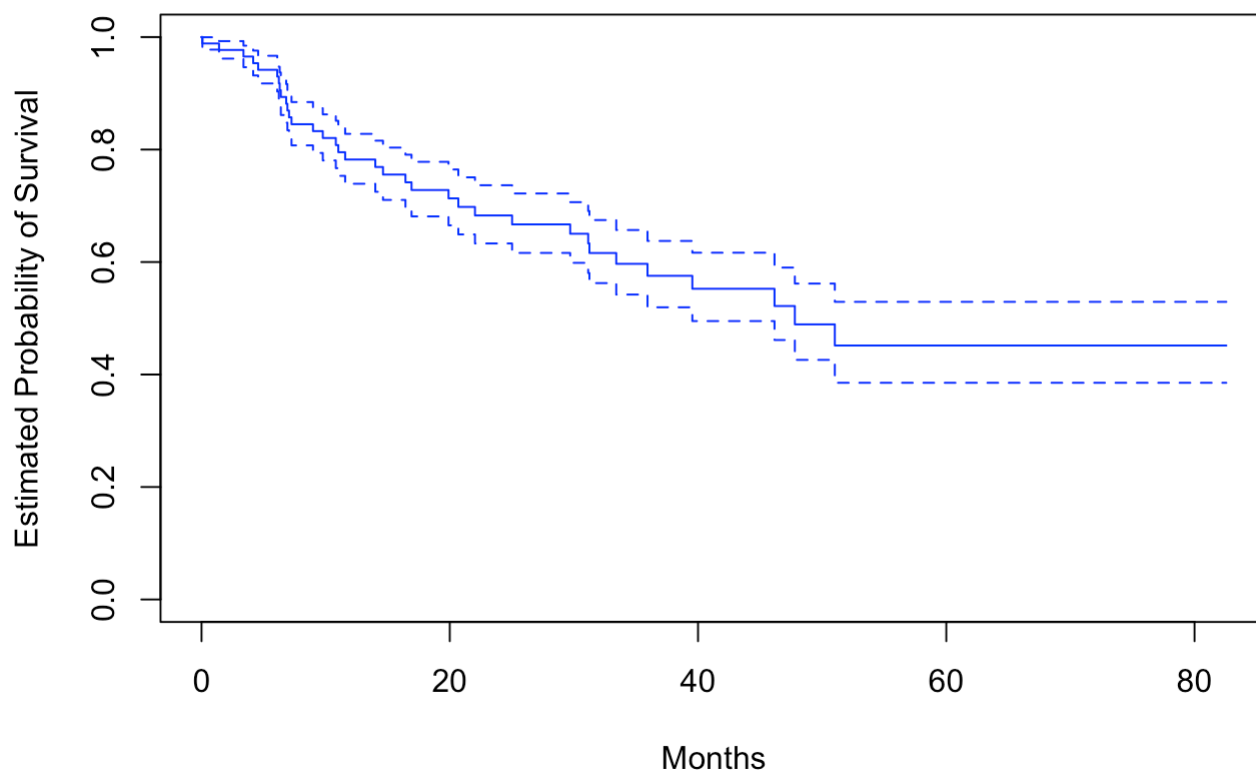
10. b.

```r
library(survival)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:nlme':
##
##     collapse
```

```
## The following objects are masked from 'package:Hmisc':
##
##     src, summarize
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```r
set.seed(123) # for reproducibility
n <- 100
time <- rexp(n, rate = 0.1)
status <- rbinom(n, size = 1, prob = 0.7)
data <- data.frame(time = time, status = status)

fit <- survfit(Surv(time, status) ~ 1)

n_bootstraps <- 200
bootstrap_results <- replicate(n_bootstraps, {
  boot_data <- sample_n(data, size = n, replace = TRUE)
  survfit(Surv(boot_data$time, boot_data$status) ~ 1)
}, simplify = FALSE)

bootstrap_surv_probs <- sapply(bootstrap_results, function(boot_result) {
  approx(boot_result$time, boot_result$surv, xout = fit$time)$y
})

se <- apply(bootstrap_surv_probs, 1, sd)

plot(fit, xlab = "Months", ylab = "Estimated Probability of Survival", col = "blue",
conf.int = 0.67)
lines(fit$time, fit$surv - se, lty = 2, col = "red")
lines(fit$time, fit$surv + se, lty = 2, col = "red")
```
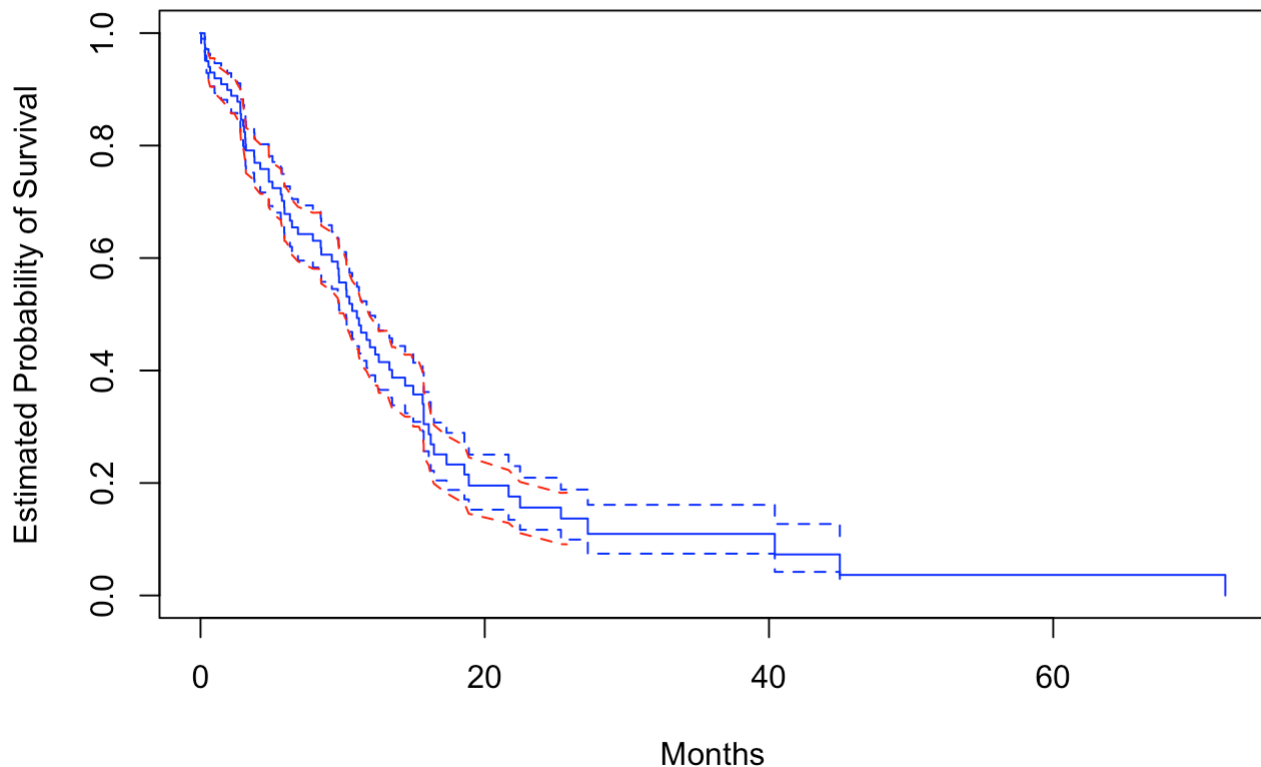
10. c.

diagnosisHG and ki seem to be significant.

```
fit <- coxph(Surv(time, status) ~ sex + diagnosis + loc + ki + gtv + stereo, data = B
rainCancer)
fit
```

```
## Call:
## coxph(formula = Surv(time, status) ~ sex + diagnosis + loc +
##     ki + gtv + stereo, data = BrainCancer)
##
##                      coef exp(coef) se(coef)      z        p
## sexMale           0.18375   1.20171  0.36036  0.510  0.61012
## diagnosisLG glioma 0.91502  2.49683  0.63816  1.434  0.15161
## diagnosisHG glioma 2.15457  8.62414  0.45052  4.782 1.73e-06
## diagnosisOther    0.88570   2.42467  0.65787  1.346  0.17821
## locSupratentorial 0.44119   1.55456  0.70367  0.627  0.53066
## ki               -0.05496   0.94653  0.01831 -3.001  0.00269
## gtv               0.03429   1.03489  0.02233  1.536  0.12466
## stereoSRT         0.17778   1.19456  0.60158  0.296  0.76760
##
## Likelihood ratio test=41.37  on 8 df, p=1.776e-06
## n= 87, number of events= 35
##    (1 observation deleted due to missingness)
```

10. d.

In order to account for additional predictors, we construct a model that includes those predictors. We then utilize this model to forecast new, synthesized data points. Here, we allow `ki` to assume every possible value, while maintaining the other predictors at either their mode or mean values.

```r
library(ggfortify)
```

```
## Warning: package 'ggfortify' was built under R version 4.3.2
```

```r
modeldata <- data.frame(
  sex = rep("Female", 5),
  diagnosis = rep("Meningioma", 5),
  loc = rep("Supratentorial", 5),
  ki = c(60, 70, 80, 90, 100),
  gtv = rep(mean(BrainCancer$gtv), 5),
  stereo = rep("SRT", 5)
)
survplots <- survfit(fit, newdata = modaldata)
plot(survplots, xlab = "Months", ylab = "Survival Probability", col = 2:6)
legend("bottomleft", c("60", "70", "80", "90", "100"), col = 2:6, lty = 1)
```