HW2

Benjamin Panny

2023-10-02

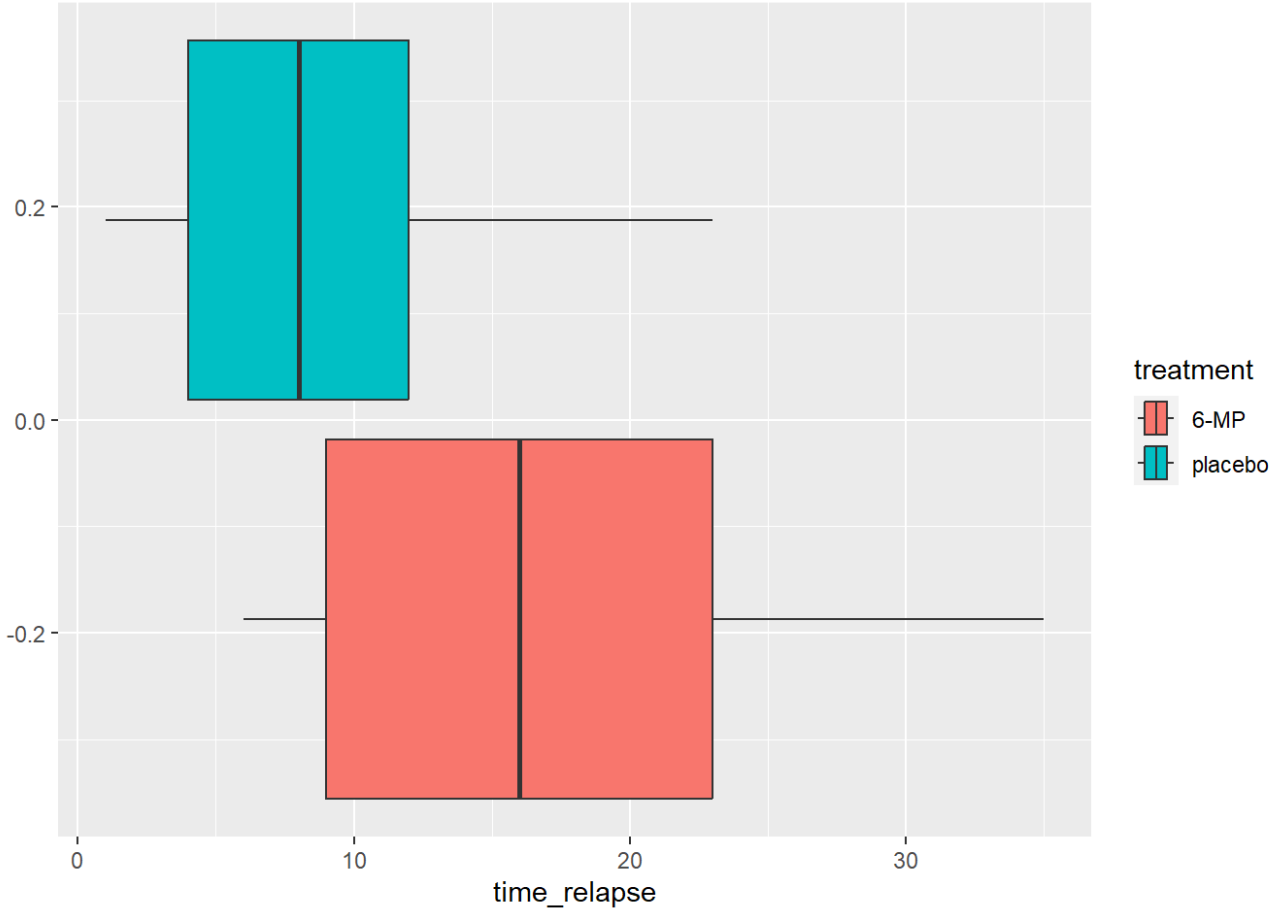
```
leuk <- read_csv('leukemia.csv')
```

```
## Rows: 42 Columns: 6
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (1): treatment
## dbl (5): ID, pair, remission_status, time_relapse, event_indicator
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

1. Leukemia Treatment Study

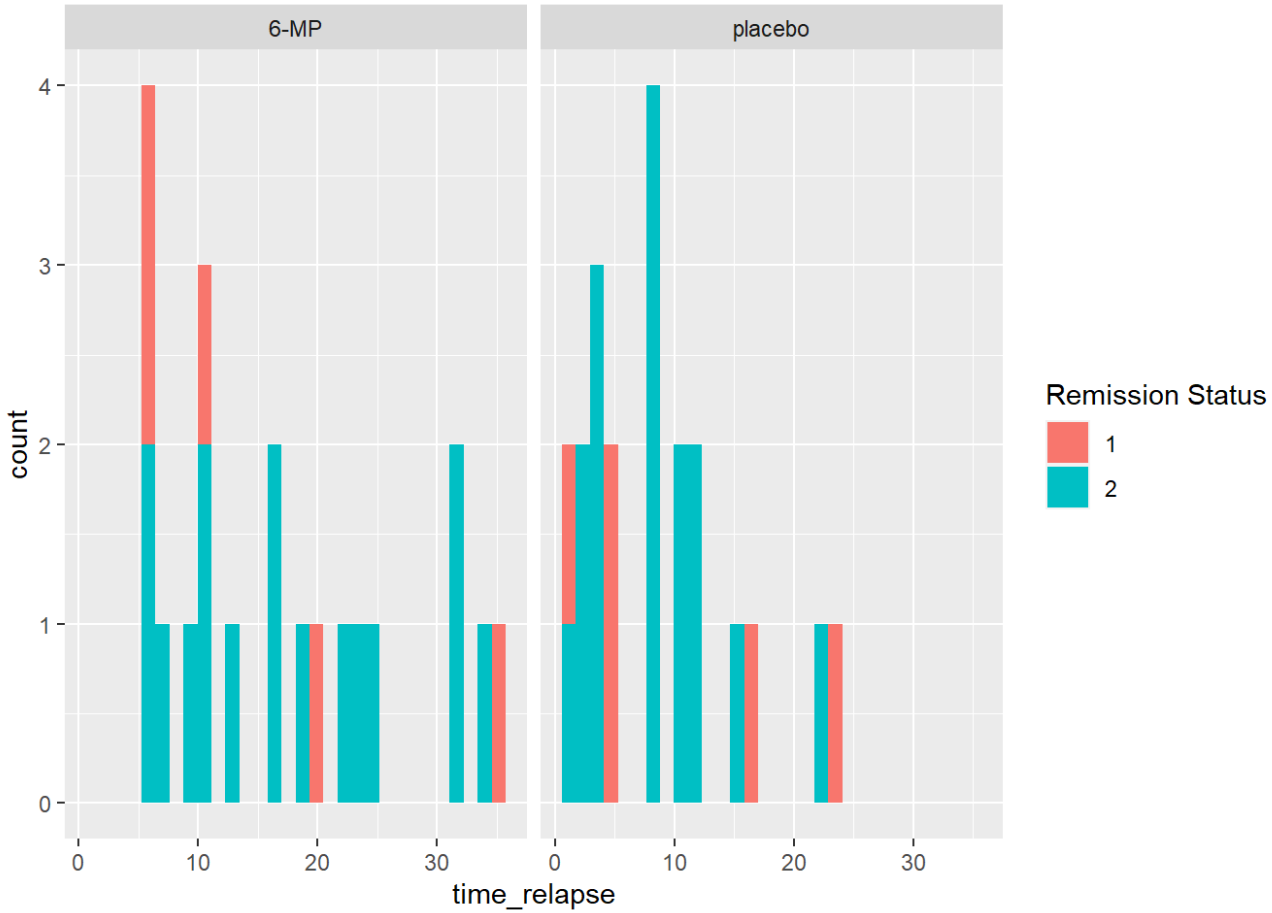
a. (10 pts) Describe what you have observed regarding “time-to-relapse” in each treatment group. (Use SAS or R) For each group, report the total number of events

```
#rem status = 1 is full remission, rem status = 2 is partial remission
#trial conducted by matching pairs of patients by remission status and randomizing within the pair to either 6-Mp or placebo maintenace
leuk %>%
  ggplot(aes(x = time_relapse, fill = treatment))+
  geom_boxplot()
```




```
leuk %>%
  ggplot(aes(x = time_relapse, fill = factor(remission_status)))+
  geom_histogram()+
  facet_wrap(~treatment)+
  labs(fill = "Remission Status")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The “time_to_relapse” distribution for the placebo group is shifted to the left compared to the 6-MP group.

```
leuk %>%
  count(treatment, event_indicator)
```

```
## # A tibble: 3 × 3
##   treatment event_indicator     n
##   <chr>               <dbl> <int>
## 1 6-MP                    0    12
## 2 6-MP                    1     9
## 3 placebo                 1    21
```

The total number of events in the 6-MP group is 9 and the total number of events in the placebo group is 21. This table indicates that everyone in the placebo group experienced an event, while not everyone in the 6-MP group did.

a. For each group, report the median time-to-relapse, and the 95% pointwise CI for the median.
b. (10 pts) Use SAS or R to generate K-M estimators for each treatment group. Plot the two K-M curves in a single plot.

```
leuk_1 <- survfit(Surv(time_relapse,event_indicator) ~ treatment, data=leuk,conf.type="log-log")
summary(leuk_1)
```

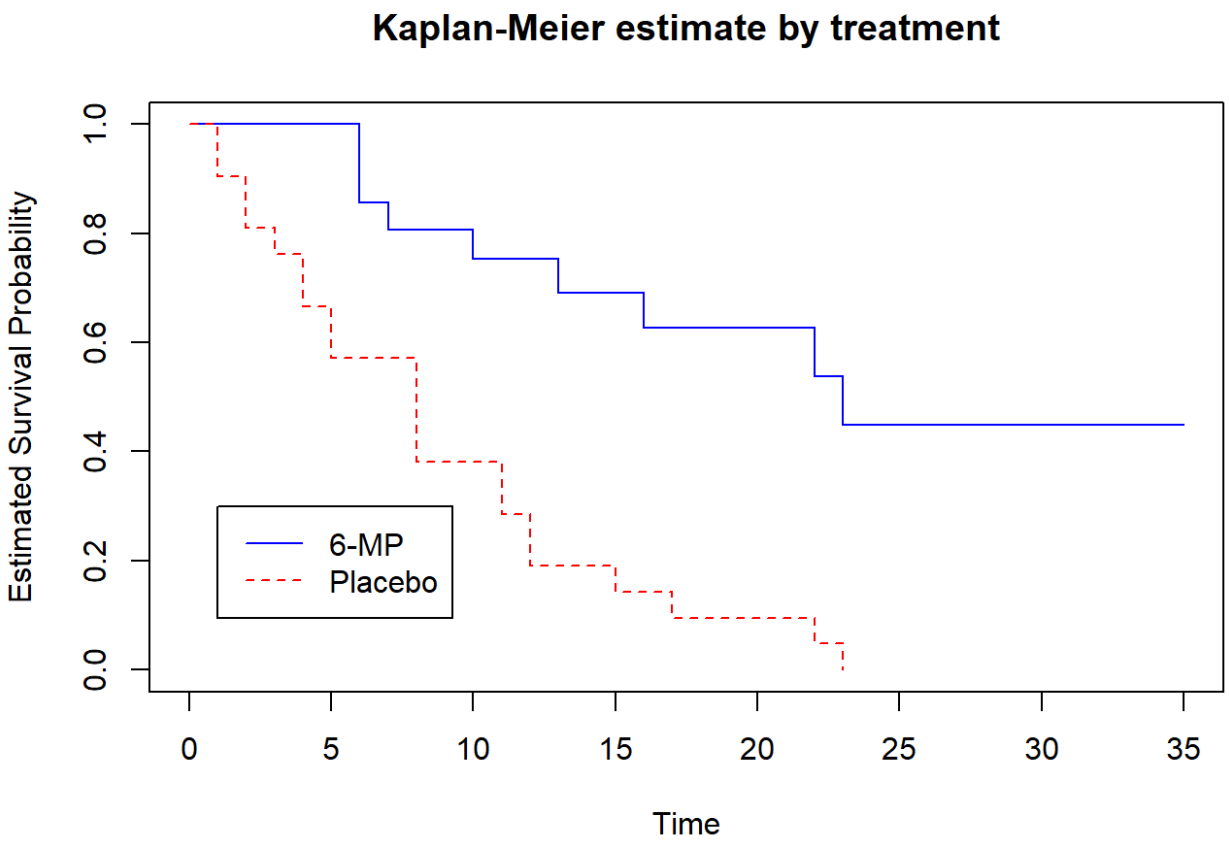
```
## Call: survfit(formula = Surv(time_relapse, event_indicator) ~ treatment,
##     data = leuk, conf.type = "log-log")
##
##                 treatment=6-MP
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     6     21       3    0.857  0.0764        0.620        0.952
##     7     17       1    0.807  0.0869        0.563        0.923
##    10     15       1    0.753  0.0963        0.503        0.889
##    13     12       1    0.690  0.1068        0.432        0.849
##    16     11       1    0.627  0.1141        0.368        0.805
##    22      7       1    0.538  0.1282        0.268        0.747
##    23      6       1    0.448  0.1346        0.188        0.680
##
##                 treatment=placebo
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1     21       2   0.9048  0.0641      0.67005        0.975
##     2     19       2   0.8095  0.0857      0.56891        0.924
##     3     17       1   0.7619  0.0929      0.51939        0.893
##     4     16       2   0.6667  0.1029      0.42535        0.825
##     5     14       2   0.5714  0.1080      0.33798        0.749
##     8     12       4   0.3810  0.1060      0.18307        0.578
##    11      8       2   0.2857  0.0986      0.11656        0.482
##    12      6       2   0.1905  0.0857      0.05948        0.377
##    15      4       1   0.1429  0.0764      0.03566        0.321
##    17      3       1   0.0952  0.0641      0.01626        0.261
##    22      2       1   0.0476  0.0465      0.00332        0.197
##    23      1       1   0.0000     NaN           NA           NA
```

```
print(leuk_1, print.rmean=TRUE)
```

```
## Call: survfit(formula = Surv(time_relapse, event_indicator) ~ treatment,
##     data = leuk, conf.type = "log-log")
##
##                     n events rmean* se(rmean) median 0.95LCL 0.95UCL
## treatment=6-MP     21      9  23.29      2.83     23      13      NA
## treatment=placebo 21     21   8.67      1.38      8       4      11
##     * restricted mean with upper limit =  35
```

The median time to relapse for the 6-MP group was 23 units of time, the lower bound of the 95% confidence interval is 13, while the upper bound could not be computed. The median time to relapse in the placebo group was 8 units of time, with a lower bound of the 95% confidence intrval at 4, and an upper bound at 11.

```
plot(leuk_1, main="Kaplan-Meier estimate by treatment", xlab="Time", ylab="Estimated Survival Probability", lty=c(1,2),col=c
("blue","red"))
legend(x=1,y=0.3,legend=c("6-MP","Placebo"),col=c("blue","red"),lty=c(1,2))
```



The survival curve for the 6-MP group appears more optimistic.

c. (10 pts) Perform the log-rank comparing the two treatment groups and interpret the result.

```
leuk_diff <- survdiff(Surv(time_relapse,event_indicator) ~ treatment, data=leuk)
leuk_diff
```

```
## Call:
## survdiff(formula = Surv(time_relapse, event_indicator) ~ treatment,
##     data = leuk)
##
##                     N Observed Expected (O-E)^2/E (O-E)^2/V
## treatment=6-MP     21        9     19.3      5.46      16.8
## treatment=placebo 21       21     10.7      9.77      16.8
##
##  Chisq= 16.8  on 1 degrees of freedom, p= 4e-05
```

The log rank test-statistic is statistically significant. This is statistical reason to reject the null hypothesis that the survival function of the 6-MP group is the same as the survival function of the placebo group for all time `t` in support of the alternative hypothesis, that the survival functions at at least some time t are unequal.

d. (10 pts) Take the paired design into account and perform a stratified log-rank test with stratification on the remission status. Output and interpret the result.

```
leuk_diff_strat <- survdiff(Surv(time_relapse,event_indicator) ~ treatment + strata(remission_status), data=leuk)
leuk_diff_strat
```

```
## Call:
## survdiff(formula = Surv(time_relapse, event_indicator) ~ treatment +
##     strata(remission_status), data = leuk)
##
##                     N Observed Expected (O-E)^2/E (O-E)^2/V
## treatment=6-MP     21        9     19.3      5.51      17.9
## treatment=placebo 21       21     10.7      9.96      17.9
##
##  Chisq= 17.9  on 1 degrees of freedom, p= 2e-05
```

The null hypothesis for the stratified log-rank test is that the survival function at time t for each stratum in each treatment group is the same for all times `t`. Our chi-square statistic is significantly unlikely given the null hypothesis, indicating a statistical reason to reject this null hypothesis in support of the alternative hypothesis, that at least one of the survival functions at at least some time t is different from the rest.

2. German Breast Cancer Study

```
gbcs <- read_tsv('gbcs.txt')
```

```
## Rows: 686 Columns: 16
## ── Column specification ────────────────────────────────────────────────────
## Delimiter: "\t"
## chr  (3): diagdate, recdate, deathdate
## dbl (13): id, age, menopause, hormone, size, grade, nodes, prog_recp, estrg_...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

diagdate = date of diagnosis recdate = date of recurrence free survival deathdate = date of death age = age at diagnosis menopause = menopausal status hormone = hormone therapy

a. (10 pts) Fit the proportional hazards model containing age, menopause status, hormone therapy, tumor size, number of nodes, number of progesterone receptors, and number of estrogen receptors. Report the model fit summary.

```
# values of menopause and hormone are 1 and 2. I'm assuming 2 == 1 and 1 == 0 in the data dictionary
gbcs_for_mod <- gbcs %>%
  mutate(menopause = factor(if_else(menopause == 2, "Yes", "No")),
         hormone = factor(if_else(hormone == 2, "Yes", "No")))
gbcs_1 <- gbcs_for_mod %>%
  coxph(Surv(rectime, censrec)~ age
        + menopause + hormone + size + nodes +
          prog_recp + estrg_recp,
        data = .)
gbcs_1_summary <- gbcs_1 %>% summary()
```

```
gbcs_1_summary
```

```
## Call:
## coxph(formula = Surv(rectime, censrec) ~ age + menopause + hormone +
##     size + nodes + prog_recp + estrg_recp, data = .)
##
##   n= 686, number of events= 299
##
##                     coef  exp(coef)   se(coef)      z Pr(>|z|)
## age          -0.0104836  0.9895712  0.0092811 -1.130  0.25866
## menopauseYes  0.2767388  1.3188219  0.1821740  1.519  0.12874
## hormoneYes   -0.3642573  0.6947124  0.1283885 -2.837  0.00455 **
## size          0.0083534  1.0083883  0.0039453  2.117  0.03423 *
## nodes         0.0498361  1.0510988  0.0074020  6.733 1.66e-11 ***
## prog_recp    -0.0026007  0.9974027  0.0005841 -4.452 8.50e-06 ***
## estrg_recp    0.0001774  1.0001774  0.0004617  0.384  0.70077
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##              exp(coef) exp(-coef) lower .95 upper .95
## age             0.9896     1.0105    0.9717    1.0077
## menopauseYes    1.3188     0.7583    0.9228    1.8847
## hormoneYes      0.6947     1.4394    0.5402    0.8935
## size            1.0084     0.9917    1.0006    1.0162
## nodes           1.0511     0.9514    1.0360    1.0665
## prog_recp       0.9974     1.0026    0.9963    0.9985
## estrg_recp      1.0002     0.9998    0.9993    1.0011
##
## Concordance= 0.687  (se = 0.015 )
## Likelihood ratio test= 94.89  on 7 df,   p=<2e-16
## Wald test            = 109  on 7 df,   p=<2e-16
## Score (logrank) test = 111.2  on 7 df,   p=<2e-16
```

b. (5 pts) Assess the significance of the model using the partial log likelihood ratio test.

The significance of the model is supported by the partial log likelihood ratio test. That is, we have statistical reason to reject the null hypothesis that all the coefficients in the model are 0.

c. (15 pts) For the model in 3(a), using the univariate Wald tests, which variables do not appear to contribute to the model? Fit a reduced model by removing the variables that are not significant from the univariate Wald test.

```
gbcs_1_summary
```

```
## Call:
## coxph(formula = Surv(rectime, censrec) ~ age + menopause + hormone +
##     size + nodes + prog_recp + estrg_recp, data = .)
##
##   n= 686, number of events= 299
##
##                     coef  exp(coef)   se(coef)      z Pr(>|z|)
## age          -0.0104836  0.9895712  0.0092811 -1.130  0.25866
## menopauseYes  0.2767388  1.3188219  0.1821740  1.519  0.12874
## hormoneYes   -0.3642573  0.6947124  0.1283885 -2.837  0.00455 **
## size          0.0083534  1.0083883  0.0039453  2.117  0.03423 *
## nodes         0.0498361  1.0510988  0.0074020  6.733 1.66e-11 ***
## prog_recp    -0.0026007  0.9974027  0.0005841 -4.452 8.50e-06 ***
## estrg_recp    0.0001774  1.0001774  0.0004617  0.384  0.70077
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##              exp(coef) exp(-coef) lower .95 upper .95
## age             0.9896     1.0105    0.9717    1.0077
## menopauseYes    1.3188     0.7583    0.9228    1.8847
## hormoneYes      0.6947     1.4394    0.5402    0.8935
## size            1.0084     0.9917    1.0006    1.0162
## nodes           1.0511     0.9514    1.0360    1.0665
## prog_recp       0.9974     1.0026    0.9963    0.9985
## estrg_recp      1.0002     0.9998    0.9993    1.0011
##
## Concordance= 0.687  (se = 0.015 )
## Likelihood ratio test= 94.89  on 7 df,   p=<2e-16
## Wald test            = 109  on 7 df,   p=<2e-16
## Score (logrank) test = 111.2  on 7 df,   p=<2e-16
```

The Wald-test for the beta coefficients is significant for hormone treatment, tumor size, number of nodes, and number of progesterone receptors. The interepretation of these exp(coefficients) is that they are hazard ratios, given unit increases in the covariates. For example, holding all other covariates constant, a patient receiving hormone treatment has ~30% lower hazard than those who are not receiving hormone treatment. As a continuous example, holding all other covariates constant, a one millimeter increase in tumor size is associated with a ~0.84% higher hazard. It appears that age, menopause status, and the number of estrogen receptors do not contribute to the model significantly.

```
gbcs_2 <- gbcs_for_mod %>%
    coxph(Surv(rectime, censrec)~ hormone + size + nodes + prog_recp,
          data = .)
gbcs_2_summary <- gbcs_2 %>% summary()
gbcs_2_summary
```

```
## Call:
## coxph(formula = Surv(rectime, censrec) ~ hormone + size + nodes +
##     prog_recp, data = .)
##
##   n= 686, number of events= 299
##
##                   coef  exp(coef)   se(coef)      z Pr(>|z|)
## hormoneYes -0.3424641  0.7100186  0.1251055 -2.737  0.00619 **
## size        0.0079221  1.0079536  0.0038971  2.033  0.04207 *
## nodes       0.0500163  1.0512882  0.0074030  6.756 1.42e-11 ***
## prog_recp  -0.0026158  0.9973876  0.0005655 -4.626 3.73e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##            exp(coef) exp(-coef) lower .95 upper .95
## hormoneYes    0.7100     1.4084    0.5556    0.9073
## size          1.0080     0.9921    1.0003    1.0157
## nodes         1.0513     0.9512    1.0361    1.0667
## prog_recp     0.9974     1.0026    0.9963    0.9985
##
## Concordance= 0.684  (se = 0.015 )
## Likelihood ratio test= 92.41  on 4 df,   p=<2e-16
## Wald test            = 106.1  on 4 df,   p=<2e-16
## Score (logrank) test = 108.9  on 4 df,   p=<2e-16
```

Then test for the significance of the variables removed using the partial log-likelihood ratio test between the full model and the reduced model.

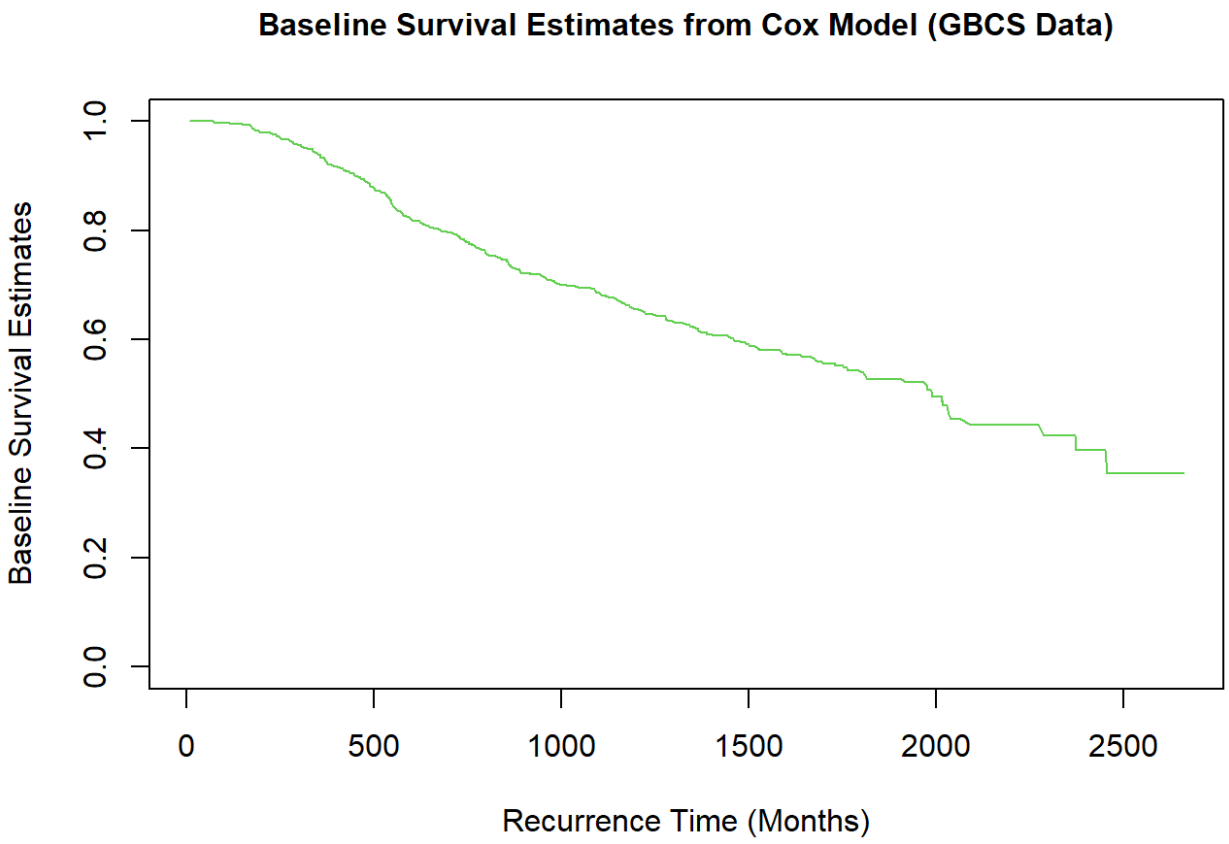
```
anova(gbcs_2,gbcs_1)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(rectime, censrec)
##  Model 1: ~ hormone + size + nodes + prog_recp
##  Model 2: ~ age + menopause + hormone + size + nodes + prog_recp + estrg_recp
##    loglik  Chisq Df Pr(>|Chi|)
## 1 -1741.9
## 2 -1740.7 2.4826  3     0.4784
```

There is no significant improvement in the model's log likelihood when using the full model compared to the reduced model.

d. (15 pts) Estimate the baseline survival function for the reduced model in problem 3(c). Plot the estimated baseline survival function (as a step function) versus recurrence time.

```
test.nc<-basehaz(gbcs_2,centered=F)
plot(test.nc$time,exp(-test.nc$hazard),type="l",lty=1,col=3,ylim=c(0,1),ylab=c("Baseline Survival Estimates"),xlab=c("Recurr
ence Time (Months)")) # no hormone; no centering
title(main=list("Baseline Survival Estimates from Cox Model (GBCS Data)",cex=1))
```



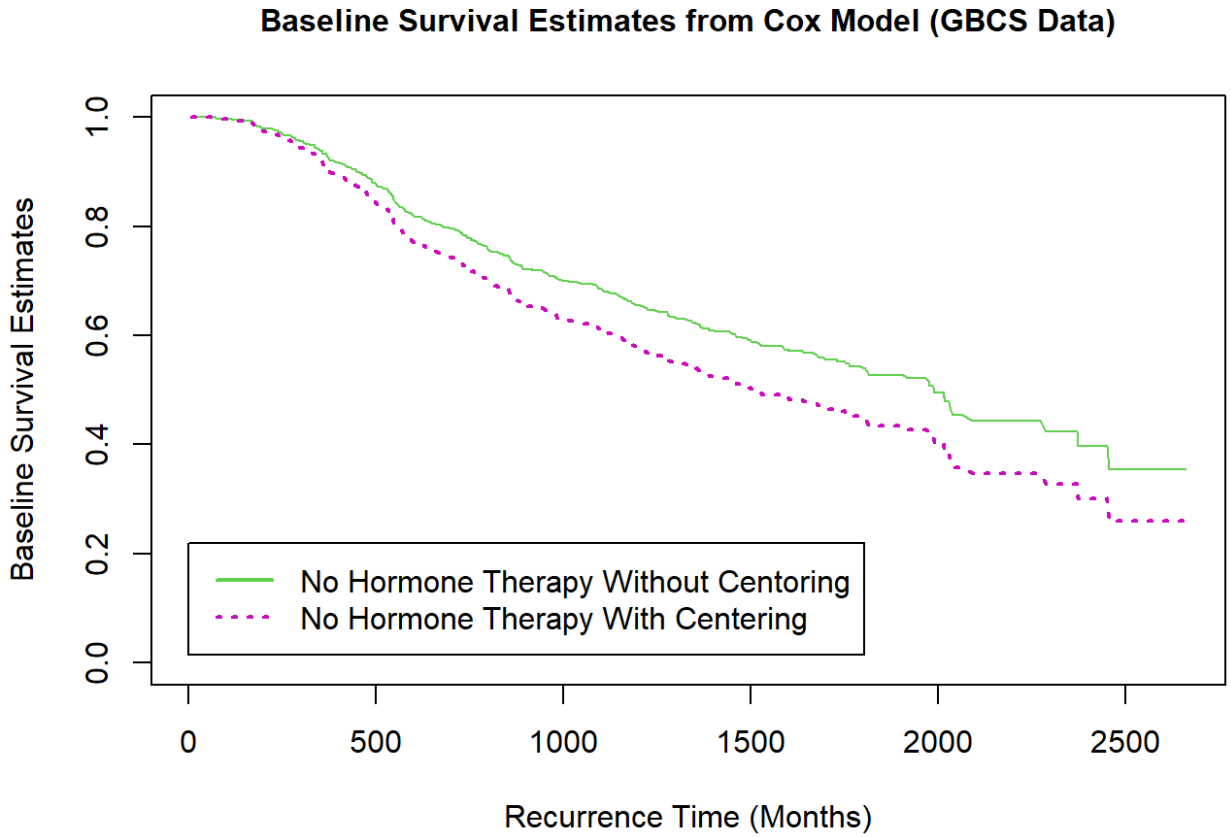
e. (15 pts) Repeat problem 1(d) fitting the model in 1(c) centering all continuous covariates at their median. Explain why the range of the estimated baseline survival functions in problems 1(d) and 1(e) are different.

```
gbcs_for_mod_3 <- gbcs_for_mod %>%
  mutate(size_c = size - median(size),
         nodes_c = nodes - median(nodes),
         prog_recp_c = prog_recp - median(prog_recp))
gbcs_3 <- gbcs_for_mod_3 %>%
    coxph(Surv(rectime, censrec)~ hormone + size_c + nodes_c + prog_recp_c,
          data = .)
gbcs_3_summary <- gbcs_3 %>% summary()
gbcs_3_summary
```

```
## Call:
## coxph(formula = Surv(rectime, censrec) ~ hormone + size_c + nodes_c +
##     prog_recp_c, data = .)
##
##   n= 686, number of events= 299
##
##                    coef  exp(coef)   se(coef)      z Pr(>|z|)
## hormoneYes  -0.3424641  0.7100186  0.1251055 -2.737  0.00619 **
## size_c       0.0079221  1.0079536  0.0038971  2.033  0.04207 *
## nodes_c      0.0500163  1.0512882  0.0074030  6.756 1.42e-11 ***
## prog_recp_c -0.0026158  0.9973876  0.0005655 -4.626 3.73e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##             exp(coef) exp(-coef) lower .95 upper .95
## hormoneYes     0.7100     1.4084    0.5556    0.9073
## size_c         1.0080     0.9921    1.0003    1.0157
## nodes_c        1.0513     0.9512    1.0361    1.0667
## prog_recp_c    0.9974     1.0026    0.9963    0.9985
##
## Concordance= 0.684  (se = 0.015 )
## Likelihood ratio test= 92.41  on 4 df,   p=<2e-16
## Wald test            = 106.1  on 4 df,   p=<2e-16
## Score (logrank) test = 108.9  on 4 df,   p=<2e-16
```

```
test.c<-basehaz(gbcs_3,centered=F)

plot(test.nc$time,exp(-test.nc$hazard),type="l",lty=1,col=3,ylim=c(0,1),ylab=c("Baseline Survival Estimates"),xlab=c("Recurr
ence Time (Months)")) # no hormone; no centering
lines(test.c$time,exp(-test.c$hazard),type="l",lty=3,col=6,lwd=2) # no hormone; centering
title(main=list("Baseline Survival Estimates from Cox Model (GBCS Data)",cex=1))
legend(x=0,y=0.22,c("No Hormone Therapy Without Centoring","No Hormone Therapy With Centering"),lty=c(1,3),col=c(3,6),lwd=2)
```




```
gbcs_summary <- gbcs  %>%
  summarize(size_median = median(size),
         nodes_median = median(nodes),
         prog_recp_median = median(prog_recp))
```

The range of the estimated baseline survival functions in problems 2(d) and 2(e) are different because 2d) is the “baseline” survival function, where `hormone treatment = None` , `tumor size = 0mm` , `# nodes = 0` , `# prog_recp = 0` , while 2e) is the “baseline/median” survival function, where `hormone treatment = None` , tumor size = 25, # nodes = 3, # prog_recp = 32.5