hw5

Benjamin Panny

2023-12-04

1. (20 pts + 10 extra pts) The model building example in Chapter 5 involved finding the best model in the WHAS500 data for lenfol (in years) as survival time and fstat as the censoring variable. The fit and adherence to model assumptions were assessed in Chapter 6. In this problem, we call this final model the “WHAS500 model.” (Lecture 12, Slide 26, “preliminary final model’).

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.2     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.2     ✓ tibble    3.2.1
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(survival)
whas <- read_delim("whas500.txt", delim = " ")
```

```
## Rows: 500 Columns: 22
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: " "
## chr  (3): admitdate, disdate, fdate
## dbl (19): id, age, sex, hr, sysbp, diasbp, bmi, cvd, afb, sho, chf, av3, mio...
##
## 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.
```

1a. (10 pts) Treating cohort year (variable: year) as a stratification variable, refit the WHAS500 model. Compare the estimated coefficients from this stratified model with those from the fit of the WHAS500 model. Are there any important differences (i.e., changes greater than 15 percent)?

```
whas %>% glimpse
```

```
## Rows: 500
## Columns: 22
## $ id        <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 1…
## $ age       <dbl> 83, 49, 70, 70, 70, 70, 57, 55, 88, 54, 48, 75, 48, 54, 67, …
## $ sex       <dbl> 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, …
## $ hr        <dbl> 89, 84, 83, 65, 63, 76, 73, 91, 63, 104, 95, 154, 85, 95, 93…
## $ sysbp     <dbl> 152, 120, 147, 123, 135, 83, 191, 147, 209, 166, 160, 193, 1…
## $ diasbp    <dbl> 78, 60, 88, 76, 85, 54, 116, 95, 100, 106, 110, 123, 80, 65,…
## $ bmi       <dbl> 25.54051, 24.02398, 22.14290, 26.63187, 24.41255, 23.24236, …
## $ cvd       <dbl> 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 0, 1, 0, 0, 1, …
## $ afb       <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, …
## $ sho       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
## $ chf       <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 1, …
## $ av3       <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, …
## $ miord     <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, …
## $ mitype    <dbl> 0, 1, 1, 1, 1, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, …
## $ year      <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
## $ admitdate <chr> "01/13/1997", "01/19/1997", "01/01/1997", "02/17/1997", "03/…
## $ disdate   <chr> "01/18/1997", "01/24/1997", "01/06/1997", "02/27/1997", "03/…
## $ fdate     <chr> "12/31/2002", "12/31/2002", "12/31/2002", "12/11/1997", "12/…
## $ los       <dbl> 5, 5, 5, 10, 6, 1, 5, 4, 4, 5, 5, 10, 7, 21, 4, 1, 13, 14, 6…
## $ dstat     <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
## $ lenfol    <dbl> 2178, 2172, 2190, 297, 2131, 1, 2122, 1496, 920, 2175, 2173,…
## $ fstat     <dbl> 0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 1, …
```

```
lecture_mod <- coxph(Surv(lenfol,fstat)~I((bmi/10)^2) + I((bmi/10)^3) + age*sex + hr + diasbp + chf, data=wha
s)

strat_mod <- coxph(Surv(lenfol,fstat)~I((bmi/10)^2) + I((bmi/10)^3) + age*sex + hr + diasbp + chf + strata(yea
r), data=whas)
```

bmifp1 and bmifp2 in the lecture model are annotated as bmi^2 and bmi^3, respectively, but in the lecture code it is the same with each divided by 10, so I will use the lecture code transformation.

```
whas_coefs <- lecture_mod$coefficients %>%
  rbind(strat_mod$coefficients)

(whas_coef_diffs <- (whas_coefs[1,] - whas_coefs[2,]) / whas_coefs[1,] * 100)
```

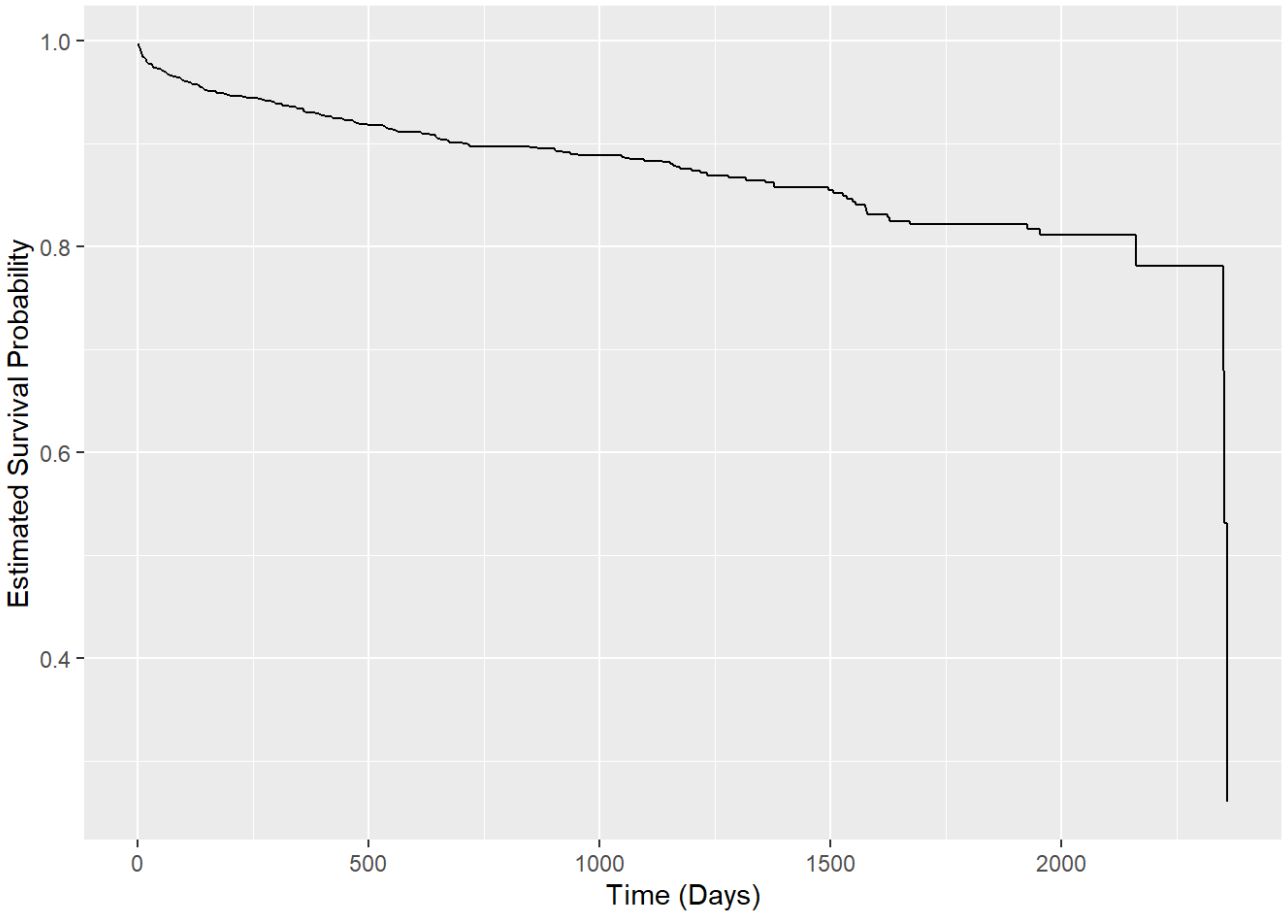
```
## I((bmi/10)^2) I((bmi/10)^3)           age           sex            hr
##   -0.73655476    0.99300717   -0.09002791  -10.62334146   10.95898580
##        diasbp           chf       age:sex
##   -2.33697238    0.14500093  -12.30347359
```

There are no changes in coefficients greater than 15% when stratifying by year. However, chf and age:sex intergtsgic come *very* close, within ~0.5% and ~3.7%, respectively.

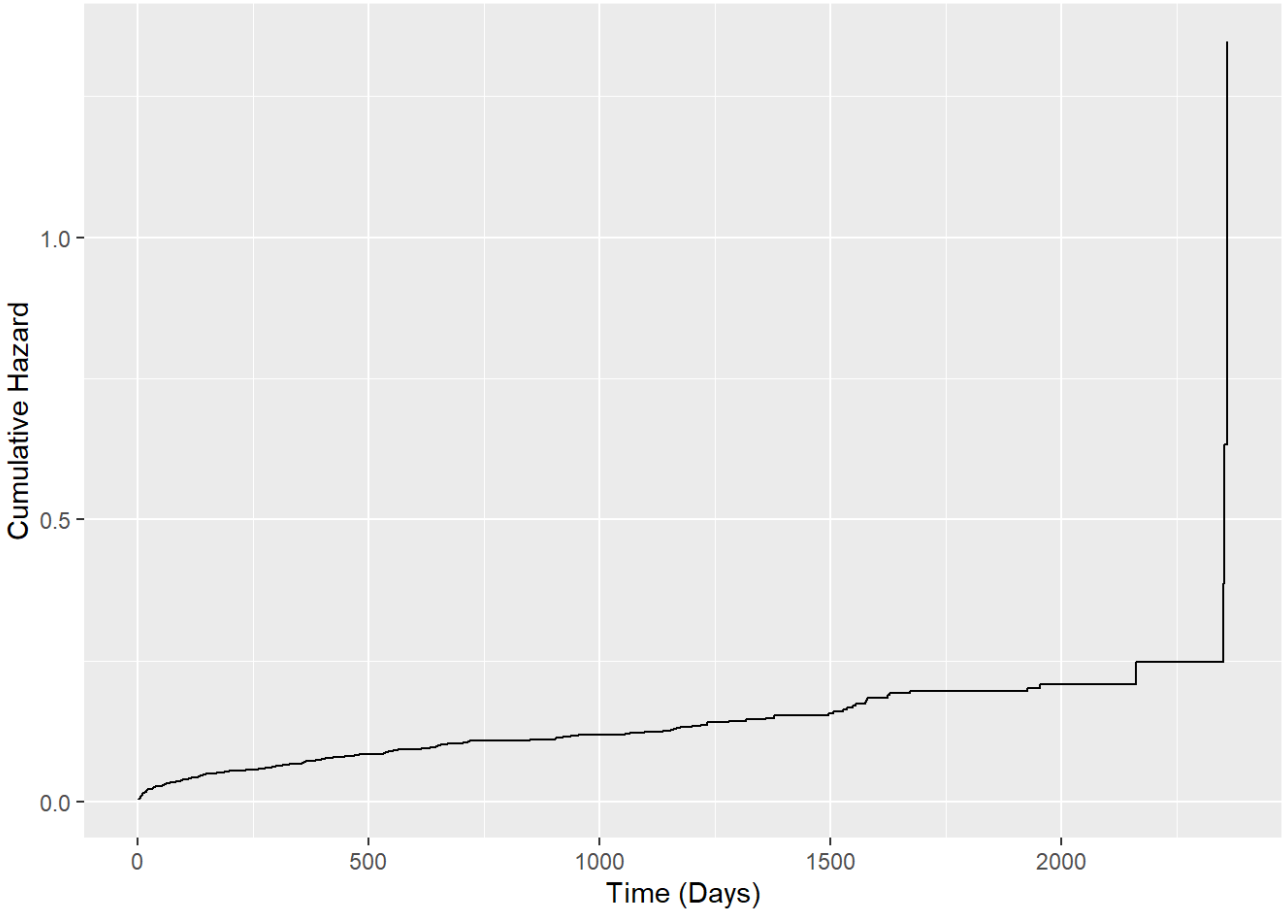
1b. (10 pts) Under the “WHAS500 model”, obtain and plot the estimated survival curve for the indivual with the following covariates:

bmi = 28, age = 50, hr = 80, diasbp = 100, chf = 1, sex = 0 (male).

```
#H0: cumulative baseline hazards for both strata
ind_covs <- data.frame(bmi = 28, age = 50, sex = 0, hr = 80, diasbp = 100, chf = 1)
H0 <- basehaz(lecture_mod,newdata = ind_covs)
S0 <- exp(-H0$hazard)

tibble(s0 = S0, time = H0$time) %>%
  ggplot(aes(x = time, y = s0)) +
  geom_step() +
  labs(x = "Time (Days)", y = "Estimated Survival Probability")
```




```
as_tibble(H0) %>%
  ggplot(aes(x = time, y = hazard)) +
  geom_step() +
  labs(x = "Time (Days)", y = "Cumulative Hazard")
```

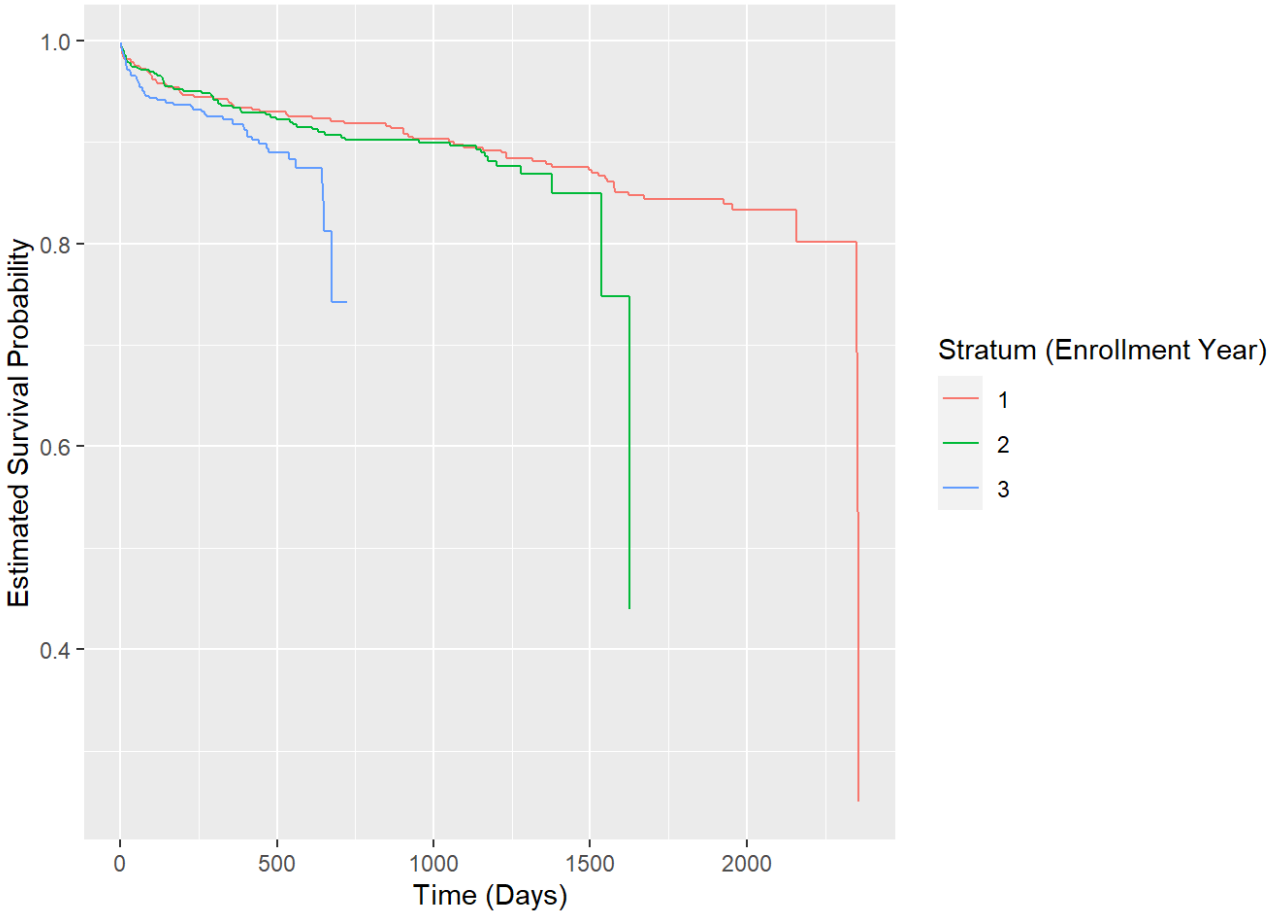


The sudden drop (survival) and rise (hazard) in the plots is surprising.

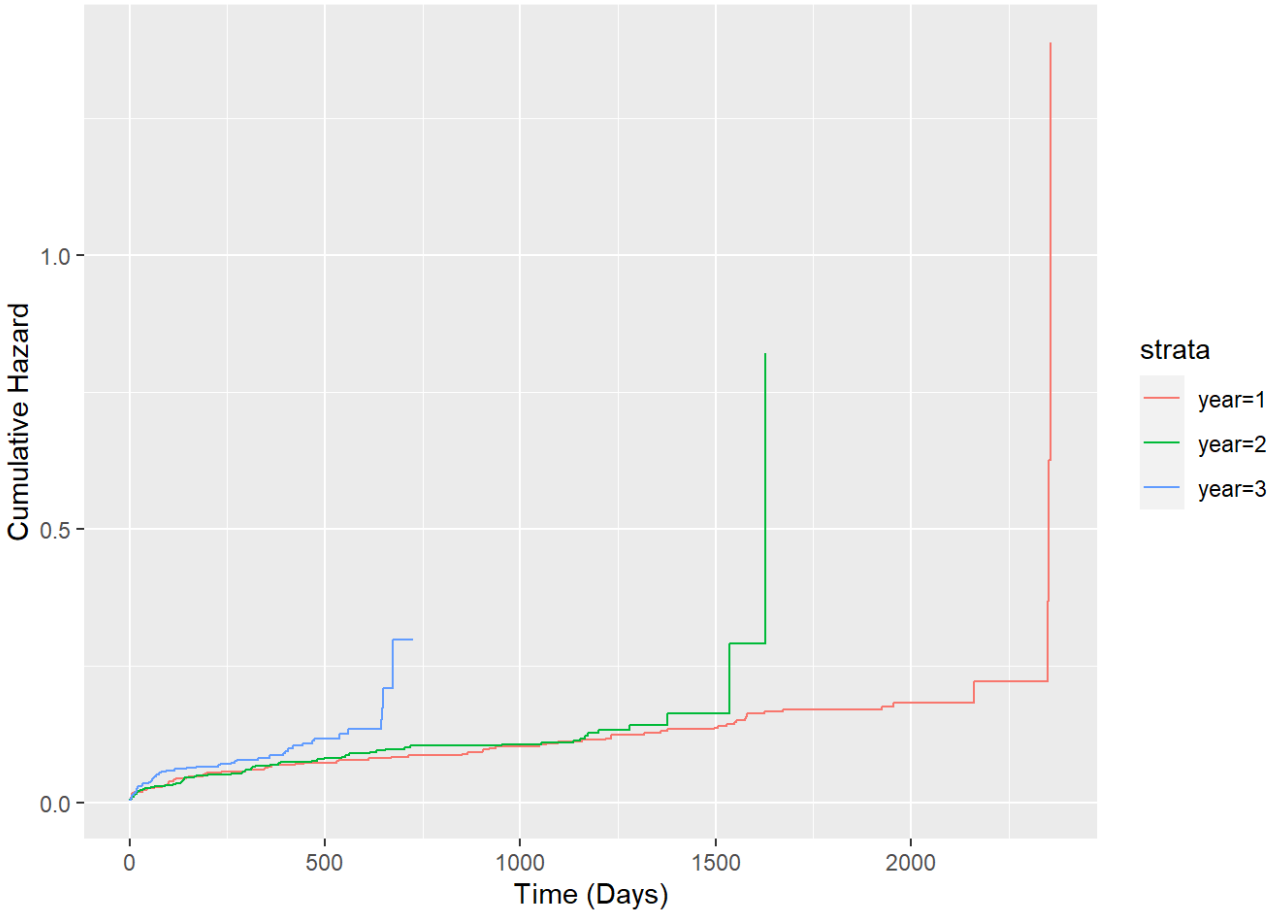
1c. (10 extra pts) Under the stratified model you fit in (a), obtain and plot the estimated survival curves for the following three individuals who have the same following covariate values (same as in (b)) but entered the study in three different cohort years (value of year=1, 2, 3, respectively):

bmi = 28, age = 50, hr = 80, diasbp = 100, chf = 1, sex = 0 (male).

```
#H0_strat: cumulative baseline hazards for both strata
H0_strat <- basehaz(strat_mod,newdata=ind_covs)
S0_strat <- exp(-H0_strat$hazard)

# cumulative baseline hazard for stratum year=1
H0_strat1 <- H0_strat[H0_strat$strata=="year=1",]
S0_strat1 <- exp(-H0_strat1$hazard)

# cumulative baseline hazard for stratum year=2
H0_strat2 <- H0_strat[H0_strat$strata=="year=2",]
S0_strat2 <- exp(-H0_strat2$hazard)

# cumulative baseline hazard for stratum year=3
H0_strat3 <- H0_strat[H0_strat$strata=="year=3",]
S0_strat3 <- exp(-H0_strat3$hazard)

tibble(s0 = c(S0_strat1, S0_strat2, S0_strat3), time = c(H0_strat1$time, H0_strat2$time, H0_strat3$time), stra
t = rep(c(1,2,3), c(length(S0_strat1), length(S0_strat2), length(S0_strat3)))) %>%
  ggplot(aes(x = time, y = s0, color = factor(strat))) +
  geom_step() +
  labs(x = "Time (Days)", y = "Estimated Survival Probability", color = "Stratum (Enrollment Year)")
```




```
as_tibble(rbind(H0_strat1, H0_strat2, H0_strat3)) %>%
  ggplot(aes(x = time, y = hazard, color = strata)) +
  geom_step() +
  labs(x = "Time (Days)", y = "Cumulative Hazard")
```



Each strata demonstrates a pattern of sharp decreases in survival and increases in hazard as the study approaches its end date.

2. (40 pts) In the dataset “GTSG_LONG.txt”, data from a clinical trial of chemotherapy and chemotherapy combined with radiotherapy in treating locally unresectable gastric cancer is given. Of interest in this study is a comparison of the efficacy of the two treatments on overall survival.

t: time to death trt: treatment, 1= chemotherapy, 2= chemotherapy combined with radiotherapy c: indicator of event, 1=event, 0=censoring

```
library(tidyverse)
library(survival)
gtsg <- read_csv('GTSG_LONG.csv')
```

```
## Rows: 90 Columns: 3
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## dbl (3): t, trt, c
##
## 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.
```

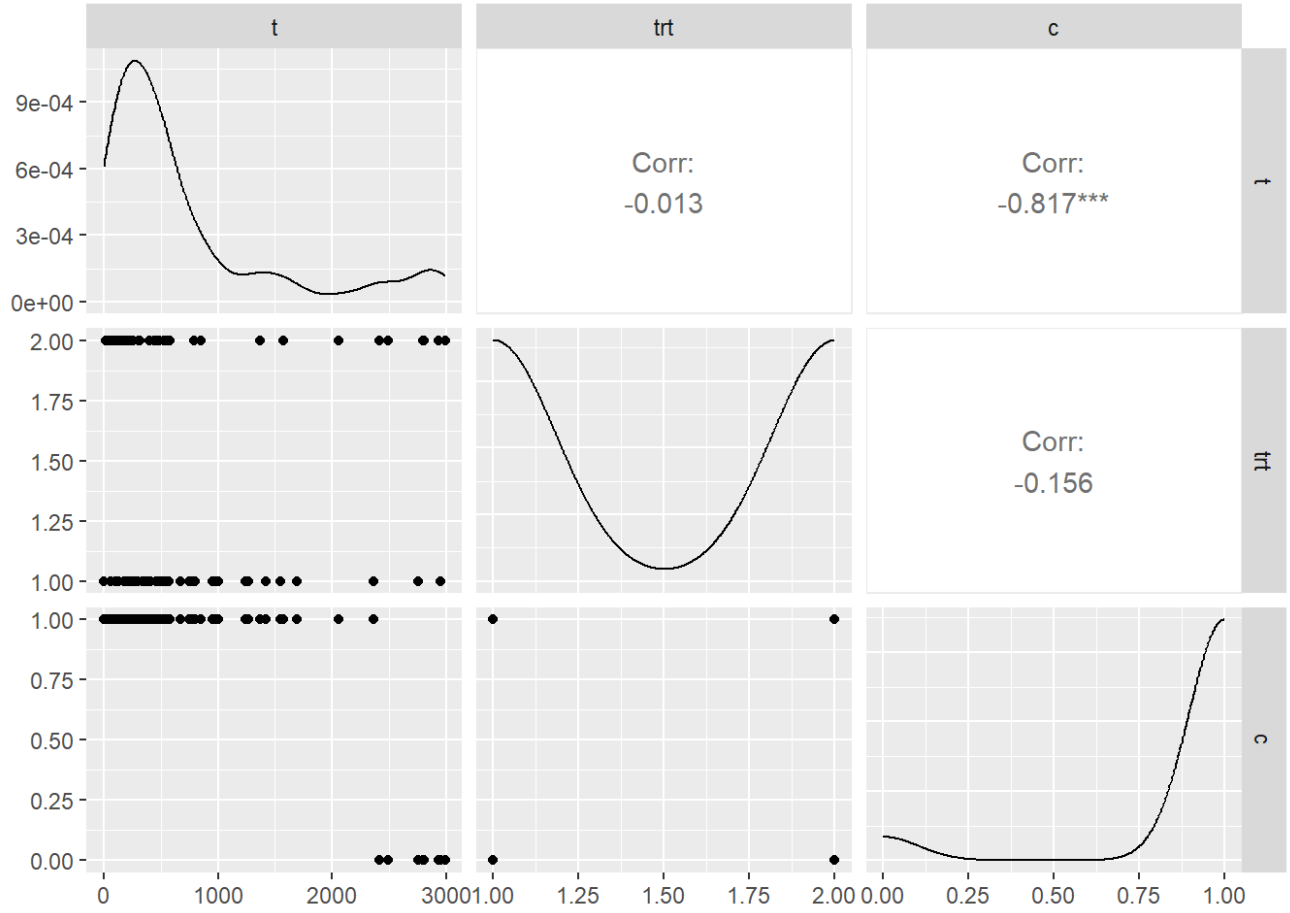
```
gtsg
```

```
## # A tibble: 90 × 3
##        t   trt     c
##    <dbl> <dbl> <dbl>
##  1     1     1     1
##  2    17     2     1
##  3    42     2     1
##  4    44     2     1
##  5    48     2     1
##  6    60     2     1
##  7    63     1     1
##  8    72     2     1
##  9    74     2     1
## 10    95     2     1
## # ℹ 80 more rows
```

2a. (10 pts) Using an appropriate proportional hazards model, test the hypothesis of difference in survival between the two treatment regimes. Find a 95% confidence interval for the hazard ratio of death for patients treated only with chemotherapy compared to patients treated with chemotherapy plus radiation.

```
gtsg %>% GGally::ggpairs()
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```




```
gtsg_mod <- coxph(Surv(t,c)~trt, data=gtsg)

summary(gtsg_mod)
```

```
## Call:
## coxph(formula = Surv(t, c) ~ trt, data = gtsg)
##
##   n= 90, number of events= 82
##
##       coef exp(coef) se(coef)     z Pr(>|z|)
## trt 0.1067    1.1126   0.2234 0.478    0.633
##
##     exp(coef) exp(-coef) lower .95 upper .95
## trt     1.113     0.8988    0.7182     1.724
##
## Concordance= 0.562  (se = 0.031 )
## Likelihood ratio test= 0.23  on 1 df,   p=0.6
## Wald test            = 0.23  on 1 df,   p=0.6
## Score (logrank) test = 0.23  on 1 df,   p=0.6
```

```
confint(gtsg_mod)
```

```
##          2.5 %    97.5 %
## trt -0.3310331 0.5444891
```

The 95% CI for the hazard ratio between death for patients treated only with chemotherapy and death for patients treated with chemotherapy plus radiation is (0.7182, 1.724). That is, if our model assumptions are correct, then reproducing our procedures will capture the true hazard ratio 95% of the time in our confidence interval. This time, our confidence interval contains 1.0, which means our resulting hazard ratio is not statistically significant under the null hypothesis that it is equal to 1.

2b. (10 pts) Test whether the proportional hazards assumption holds for the “trt” variable.

The proportional hazards assumption is, in some sense, that the effect of a covariate on a change in hazard rate for any given individual is stable over time, that is, that the covariate effect is time-independent.

```
schoenfeld_resid <- cox.zph(gtsg_mod)
par(las=F)
par(mfrow=(c(1,3)))
print(schoenfeld_resid)
```

```
##        chisq df       p
## trt     13.2  1 0.00028
## GLOBAL  13.2  1 0.00028
```

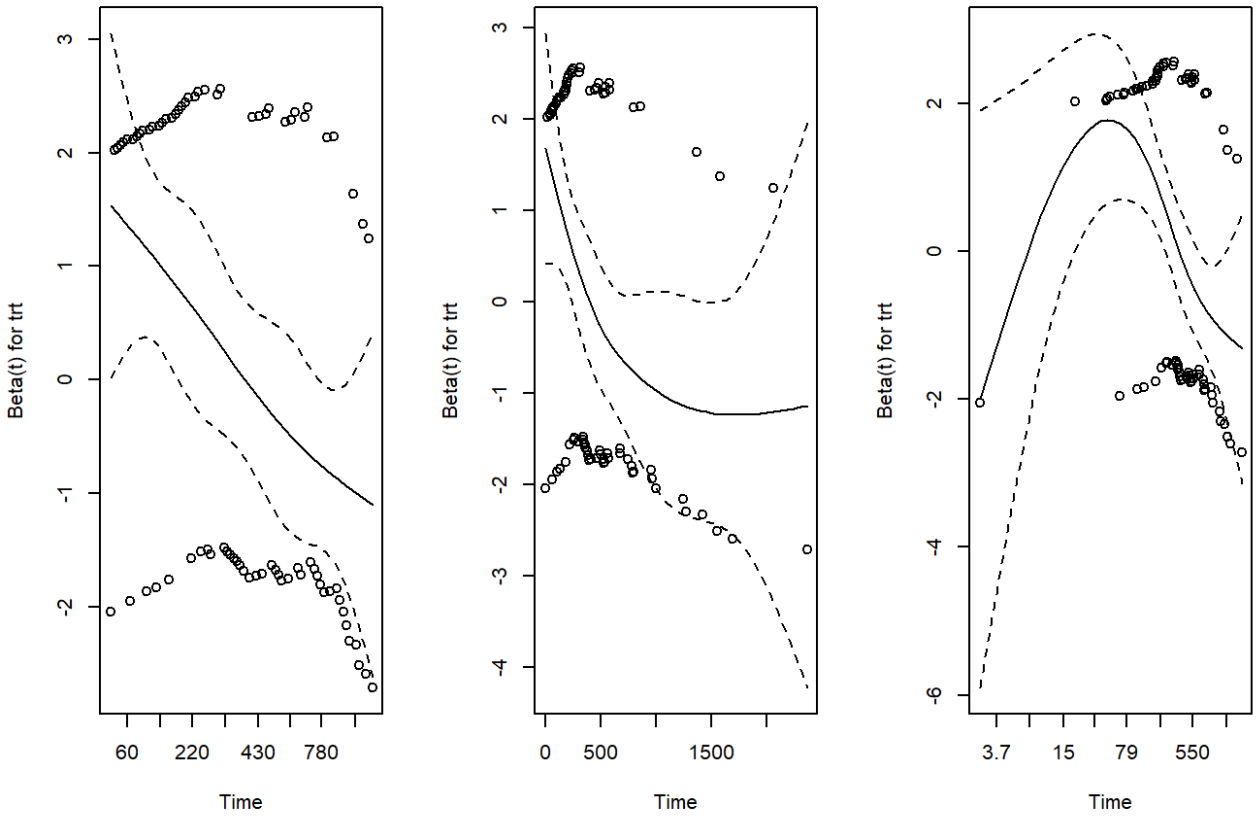
```
ss_resid_identity<-cox.zph(gtsg_mod, transform="identity")
ss_resid_identity
```

```
##        chisq df       p
## trt     10.9  1 0.00096
## GLOBAL  10.9  1 0.00096
```

```
# use the log tranformation (g(t)=log(t))
ss_resid_log<-cox.zph(gtsg_mod, transform=function(x){log(x)})
ss_resid_log
```

```
##        chisq df      p
## trt     7.35  1 0.0067
## GLOBAL  7.35  1 0.0067
```

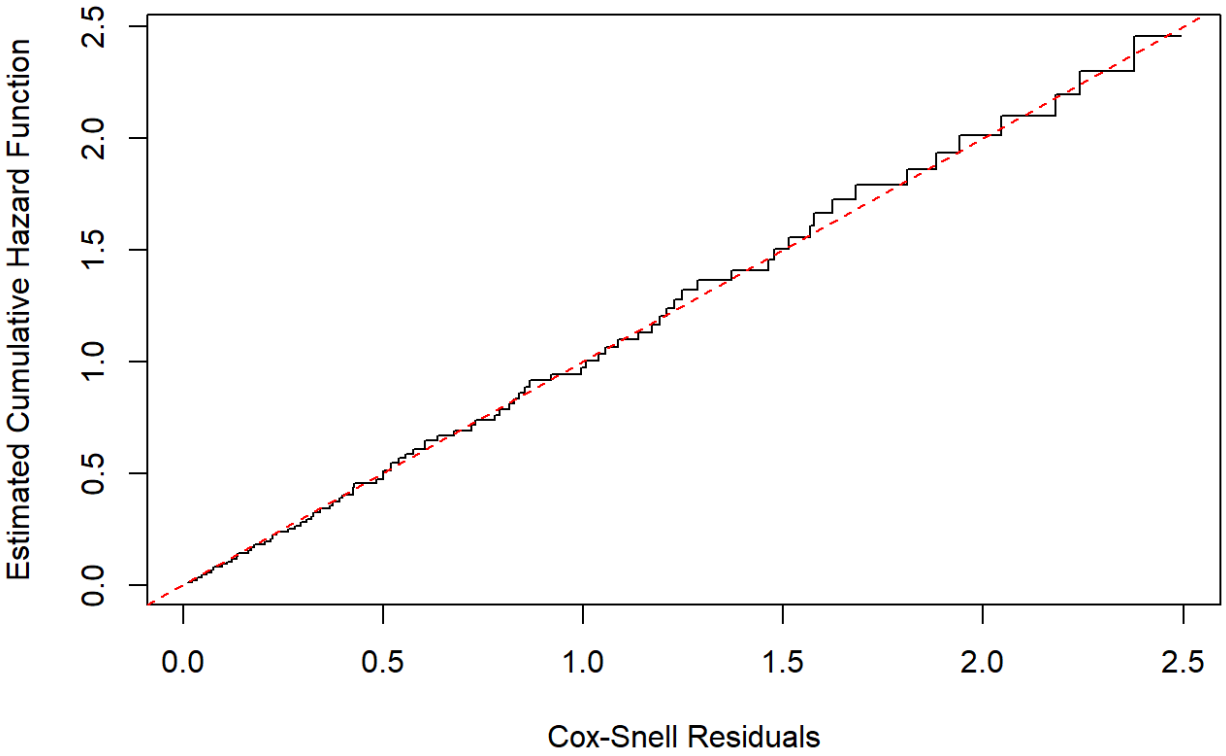
```
plot(schoenfeld_resid)
plot(ss_resid_identity)
plot(ss_resid_log)
```



The scaled Schoenfeld residuals show a time-dependent pattern under `km`, `identity`, and `log` transformations of the survival times. Therefore, the proportional hazards assumption is violated.

Model fit can also be assessed with Cox-Snell residuals.

```
coxsnell.r <- gtsg$c-resid(gtsg_mod,type="martingale")

fitres <- survfit(Surv(coxsnell.r, gtsg$c)~1)

plot(fitres$time,-log(fitres$surv),type='s',xlab='Cox-Snell Residuals',
     ylab='Estimated Cumulative Hazard Function')
abline(0,1,col='red',lty=2)
```



The model appears well fit according to the Cox-Snell Residuals because the residuals appear to match up with a unit exponential cumulative hazard function, which is what we expect to see if the Cox model is valid and the coefficients and estimated baseline hazard cumulative hazard function are close to their true values. However, as we have seen from the Scaled Schoenfeld Residuals, the Cox model is not valid!

2c. (20 pts) Because the hazard rates for the two treatment groups are not proportional (the proportional hazards assumption does not hold for “trt”), consider a model with two time-dependent covariates:

$$Z_1(t) = \begin{cases} 1 & \text{if chemotherapy only and } t \leq \tau \\ 0 & \text{otherwise} \end{cases}$$

$$Z_2(t) = \begin{cases} 1 & \text{if chemotherapy only and } t > \tau \\ 0 & \text{otherwise} \end{cases}$$

$$\text{where } \tau = 254$$

Fit the Cox PH model with these two new treatment covariates (note that you do not need the original “trt” variable in this model). Report and explain the two hazard ratios obtained from this model. Compare the result obtained from this model with the result obtained in part (a). Explain how a physician should present this model to a patient.

```
gtsg_td <- gtsg %>%
  mutate(z1 = if_else(t <= 254 & trt == 1, 1, 0),
         z2 = if_else(t > 254 & trt == 1, 1, 0))
gtsg_mod_td <- coxph(Surv(t,c)~z1 + z2, data=gtsg_td)

summary(gtsg_mod)
```

```
## Call:
## coxph(formula = Surv(t, c) ~ trt, data = gtsg)
##
##   n= 90, number of events= 82
##
##       coef exp(coef) se(coef)     z Pr(>|z|)
## trt 0.1067    1.1126   0.2234 0.478    0.633
##
##     exp(coef) exp(-coef) lower .95 upper .95
## trt     1.113     0.8988    0.7182     1.724
##
## Concordance= 0.562  (se = 0.031 )
## Likelihood ratio test= 0.23  on 1 df,   p=0.6
## Wald test            = 0.23  on 1 df,   p=0.6
## Score (logrank) test = 0.23  on 1 df,   p=0.6
```

```
summary(gtsg_mod_td)
```

```
## Call:
## coxph(formula = Surv(t, c) ~ z1 + z2, data = gtsg_td)
##
##   n= 90, number of events= 82
##
##       coef exp(coef) se(coef)      z Pr(>|z|)
## z1  1.7619    5.8235   0.4539  3.881 0.000104 ***
## z2 -0.2652    0.7670   0.2336 -1.135 0.256188
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##    exp(coef) exp(-coef) lower .95 upper .95
## z1     5.824     0.1717    2.3921    14.177
## z2     0.767     1.3037    0.4852     1.212
##
## Concordance= 0.634  (se = 0.027 )
## Likelihood ratio test= 14.25  on 2 df,   p=8e-04
## Wald test            = 19.47  on 2 df,   p=6e-05
## Score (logrank) test = 25.54  on 2 df,   p=3e-06
```

This model differs from the model obtained in part a. It indicates that there is a significance difference in hazard rate for individuals receiving chemotherapy prior to timepoint 254 compared to those receiving chemotherapy+radiation at any time point and those receiving chemotherapy who are beyond timepoint 254. The estimated hazard ratio for this difference is 5.824 (2.39, 14.177), indicating a hazard rate that is 5.824 times higher for the former compared to the latter. The hazard rate is the probability of death at the next instant of time given survival up to the current timepoint.

Additionally, this model including time-dependent covariates indicates there is no statistically significant difference in hazard rate for “individuals receiving only chemotherapy after timepoint 254” and “individuals receiving chemotherapy+radiation at anytime and individuals receiving chemotherapy prior to timepoint 254”.

My conclusion is that the physician should explain this model to the patient by indicating the chemotherapy+radiation is associated with lower hazard during the first 254 timepoints compared to chemotherapy, and afterwards there is no strong evidence that they are different. Thus, the physician might consider chemotherapy + radiation prior to 254 timepoints and then consider chemotherapy alone if it relieves burden on the patient.

3 (40 pts + 10 extra pts) Fit an exponential regression model containing Treatment (tx) and cd4 count (cd4) using the ACTG320 data (same dataset in HW4).

```
act <- read_csv('actg320.csv')
```

```
## Rows: 1151 Columns: 16
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## dbl (16): ID, TIME, CENSOR, TIME_D, CENSOR_D, TX, TXGRP, STRAT2, SEX, RACETH...
##
## 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.
```

```
names(act) <- tolower(names(act))
fit.exp.1 <- survreg(Surv(time/365.25,censor) ~ tx + cd4, dist="exp", data=act)
summary(fit.exp.1)
```

```
##
## Call:
## survreg(formula = Surv(time/365.25, censor) ~ tx + cd4, data = act,
##     dist = "exp")
##               Value Std. Error   z       p
## (Intercept) 0.81414    0.15647 5.2 2.0e-07
## tx          0.66680    0.21489 3.1  0.0019
## cd4         0.01609    0.00251 6.4 1.5e-10
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -253.5   Loglik(intercept only)= -290.1
##  Chisq= 73.36 on 2 degrees of freedom, p= 1.2e-16
## Number of Newton-Raphson Iterations: 7
## n= 1151
```

3a. (15 pts) Using the fitted model, compute point and 95 percent confidence interval estimates of the time ratio for treatment and the time ratio for an increase of 50 in cd4 Interpret these estimates within the context of the study.

```
confint(fit.exp.1)
```

```
##                  2.5 %     97.5 %
## (Intercept) 0.50747295 1.12081612
## tx          0.24562703 1.08798090
## cd4         0.01116312 0.02101472
```

```
apply(confint(fit.exp.1), FUN=exp, MARGIN=2)
```

```
##                2.5 %   97.5 %
## (Intercept) 1.661088 3.067357
## tx          1.278423 2.968275
## cd4         1.011226 1.021237
```

```
lapply(coef(fit.exp.1), FUN=exp)
```

```
## $`(Intercept)`
## [1] 2.257244
##
## $tx
## [1] 1.948001
##
## $cd4
## [1] 1.016219
```

```
print('for delta(cd4) = 50...')
```

```
## [1] "for delta(cd4) = 50..."
```

```
exp(confint(fit.exp.1)['cd4',] * 50)
```

```
##    2.5 %   97.5 %
## 1.747447 2.859755
```

```
exp(coef(fit.exp.1)['cd4'] * 50)
```

```
##      cd4
## 2.235458
```

The estimated time ratio for treatment is 1.95 (1.27, 2.97). This means the median time to event among individuals receiving treatment is 1.95 times higher than the median time to event among individuals not receiving treatment.

The estimated time ratio for an increase of 50 in cd4 is 2.23 (1.75, 2.86). This means the median time to event among individuals for each 50-unit increase in cd4 is 2.23 times higher than the median time to event at the previous cd4 level. Of course, it is unwise to extrapolate this and say, if you had infinite cd4, you would live forever.

3b. (15 pts) Using the fitted model, compute point estimates of the hazard ratio for treatment and the hazard ratio for an increase of 50 in cd4. Interpret these estimates within the context of the study. AND 3c. (10 extra pts) Compute 95 percent confidence interval estimates of the hazard ratio for treatment and the hazard ratio for an increase of 50 in cd4 (i.e., two CIs).

“Because the exponential regression model has proportional hazards, we can also express the effect of covariates using hazard ratios.” This is accomplished by changing the sign of the coefficients. This also means that the hazard ratio is the inverse time ratio and vice-versa

```
-confint(fit.exp.1)
```

```
##                   2.5 %      97.5 %
## (Intercept) -0.50747295 -1.12081612
## tx          -0.24562703 -1.08798090
## cd4         -0.01116312 -0.02101472
```

```
apply(-confint(fit.exp.1), FUN=exp, MARGIN=2)
```

```
##                 2.5 %    97.5 %
## (Intercept) 0.6020150 0.3260136
## tx          0.7822139 0.3368960
## cd4         0.9888990 0.9792046
```

```
-coef(fit.exp.1)
```

```
## (Intercept)          tx         cd4
## -0.81414453 -0.66680396 -0.01608892
```

```
lapply(-coef(fit.exp.1), FUN=exp)
```

```
## $`(Intercept)`
## [1] 0.4430182
##
## $tx
## [1] 0.5133466
##
## $cd4
## [1] 0.9840398
```

```
print('for delta(cd4) = 50...')
```

```
## [1] "for delta(cd4) = 50..."
```

```
exp(-confint(fit.exp.1)['cd4',] * 50)
```

```
##     2.5 %    97.5 %
## 0.5722634 0.3496803
```

```
exp(-coef(fit.exp.1)['cd4'] * 50)
```

```
##       cd4
## 0.4473357
```

The hazard ratio for treatment is 0.51 (0.326, 0.6). This indicates a ~49% lower hazard rate for those receiving treatment compared to those not receiving treatment

The hazard ratio for a 50 unit increase in cd4 is 0.447 (0.35, 0.57). This indicates a ~55% lower hazard rate for each 50 unit increase in cd4 compared to the previous cd4 level.

3d. (10 pts) Compare the time ratio and hazard ratio estimates computed in problems 3(a) and 3(b). In particular, which estimate time or hazard ratio would be more easily understood by non-statistically trained clinicians?

I think median survival time is far more interpretable than the hazard ratio. Indeed, the numerator and denominator of the hazard ratio are hazard rates, which are conditional instantaneous event times, with limits and infinitesimals and conditioning in the very definition. Comparatively, everyone has an intuition for what it means to say that typical length of survival is longer in one group compared to another group.