Math 150 - Methods in Biostatistics - Homework 9

Ethan Ashby

Friday, April 16, 2021

Assignment Summary (Goals)

- working with hazard functions as measures of survival (S(t) and h(t) are functions of each other!)
- working with cumulative hazard functions

Important

I put all the datasets into the Box folder (linked from Sakai), the same one which also contains the course videos. I'm hoping that having the data in Box will be easier than having it in Sakai. Note the odd format to the URL below. `/shared/static/...csv`. You should be able to read in any of the datasets using that format. In terms of this week's HW, the code I've written should read it the data just fine for you.

(Apropos of nothing, I find the following website to be very helpful in making markdown tables: https://www.tablesgenerator.com/markdown_tables)

Q1. PodQ Describe one thing you learned from someone in your pod this week (it could be: content, logistical help, background material, R information, etc.) 1-3 sentences.

Annika's Youtube video for her internship is dropping tomorrow! Gonna go viral!

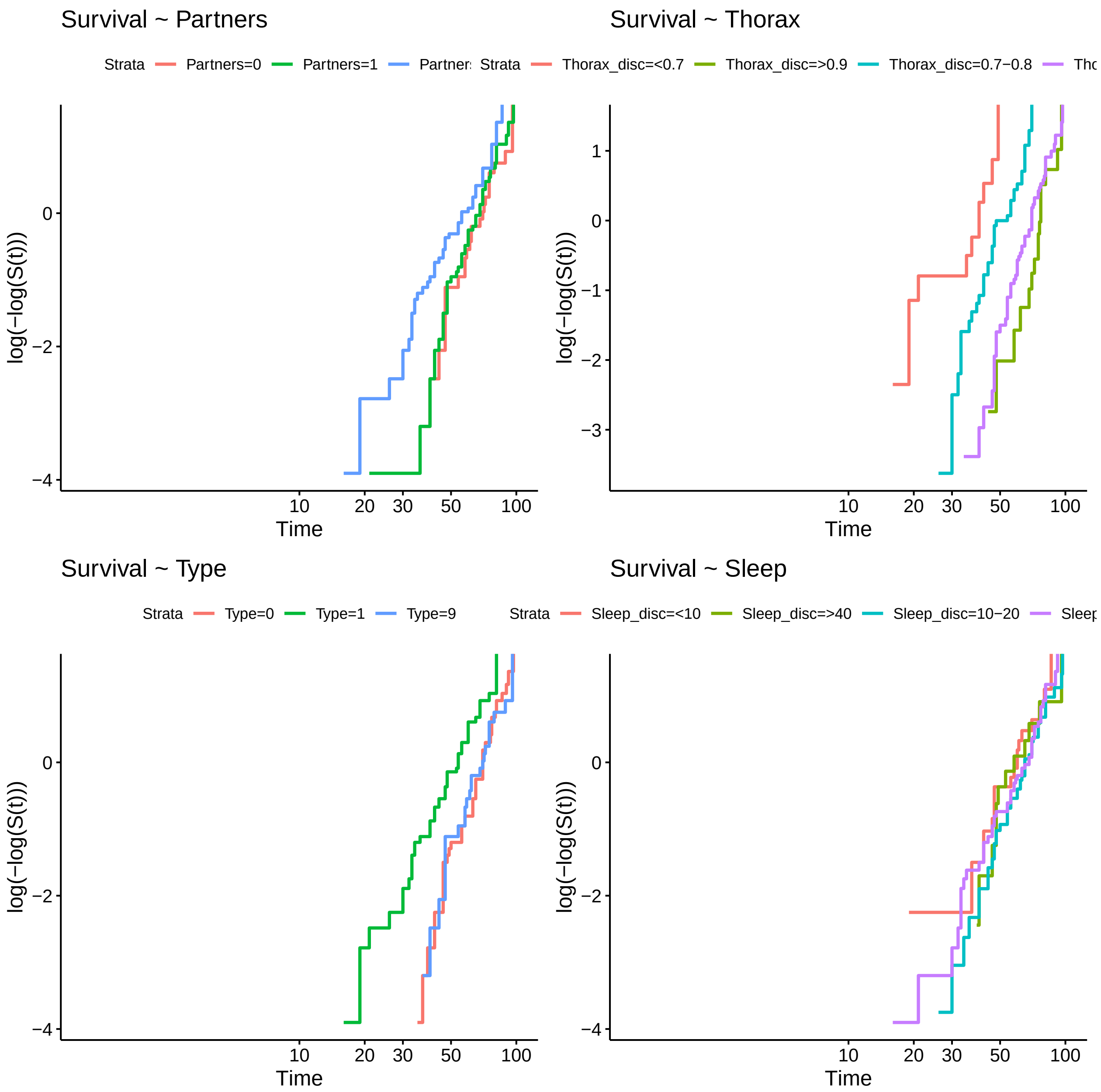
Q2. Chp 9, E11 Read text for description of data.

```
fruitfly <- read_csv("https://pomona.box.com/shared/static/qnsl0sp0twdutz6azidxb5yt37boee7v", na="*")
```

(a) check proportional hazard assumptions for treatment KM curves. Use `fun="cloglog"` inside the `ggsurvplot`.

```
surv1<-survival::survfit(Surv(Longevity, Censor)~ Partners, data=fruitfly)
surv2<-survival::survfit(Surv(Longevity, Censor)~ Type, data=fruitfly)

##########discretize continous vbls
fruitfly_edit = fruitfly %>% mutate(Thorax_disc=ifelse(Thorax<0.7, "<0.7", ifelse(between(Thorax, 0.7, 
##########

surv3<-survival::survfit(Surv(Longevity, Censor)~ Thorax_disc, data=fruitfly_edit)
surv4<-survival::survfit(Surv(Longevity, Censor)~ Sleep_disc, data=fruitfly_edit)

p1<-survminer::ggsurvplot(surv1, censor=TRUE, fun="cloglog")+ggtitle("Survival ~ Partners")
p2<-survminer::ggsurvplot(surv2, censor=TRUE, fun="cloglog")+ggtitle("Survival ~ Type")
```

```
p3<-survminer::ggsurvplot(surv3, censor=TRUE, fun="cloglog")+ggtitle("Survival ~ Thorax")
p4<-survminer::ggsurvplot(surv4, censor=TRUE, fun="cloglog")+ggtitle("Survival ~ Sleep")

arrange_ggsurvplots(list(p1, p2, p3, p4), ncol=2, nrow=2)
```



Plotting each explanatory variable (`Partners`, `Thorax`, `Type`, and `Sleep`) agains the $\log(-\log(S(t)))$ for the fruitflies largely supports the PH assumptions. Each plot shows curves that are roughly parallel; under proportional hazards, the difference between curves should be a constant vertical shift. This is mostly the case: the `Survival ~ Thorax` and `Survival~Sleep` show strong agreement with this assumption. `Survival~Partners` and `Survival~Type` show slight differences in slope between the lines, but it is likely not a major violation of the PH assumption. So we can proceed to use the Cox model with reasonable confidence that the PH assumption is satisfied.

(b) use all explanatory variables and likelihood ratio test to come up with the “best” model. [Note that the structure of the likelihood ratio test here is identical to the one we used in logistic regression! See: http://st47s.com/Math150/Notes/survival-analysis.html#multcoxph]

- After you pipe the `coxph()` model into `glanc()` you will see a column called `logLik`.
- The test statistics is $G = 2 * (logLik_{biggermodel} - logLik_{smallermodel})$
- The p-value will be calculated using a chisq distribution where the degrees of freedom are the number of extra coefficients which were estimated in the bigger model.

First, let’s fit a couple models to see if `Thorax` and `Sleep` should be encoded as continous or categorical variables:

```
coxtest1<-coxph(Surv(Longevity, Censor)~Thorax_disc, data=fruitfly_edit)
coxtest1 %>% tidy()
```

```
## # A tibble: 3 x 5
##   term                estimate std.error statistic  p.value
##   <chr>                  <dbl>     <dbl>     <dbl>    <dbl>
## 1 Thorax_disc>0.9        -3.14     0.449     -6.99 2.76e-12
## 2 Thorax_disc0.7-0.8     -1.36     0.366     -3.71 2.11e- 4
## 3 Thorax_disc0.8-0.9     -2.75     0.386     -7.14 9.39e-13
```

For the `Thorax` variable, I’m reasonably satisfied that $\ln(HR)$ is linear wrt `Thorax`, because approx 0.1 increases in the Thorax leads to approximately a -1.3 change in β (can be seen by the differences in beta between baseline (0) to 0.7-0.8 (-1.36), and 0.7-0.8 and 0.8-0.9 (-2.76)).

```
coxtest2<-coxph(Surv(Longevity, Censor)~Sleep_disc, data=fruitfly_edit)
coxtest2 %>% tidy()
```

```
## # A tibble: 3 x 5
##   term             estimate std.error statistic p.value
##   <chr>               <dbl>     <dbl>     <dbl>   <dbl>
## 1 Sleep_disc>40      -0.225     0.369    -0.608   0.543
## 2 Sleep_disc10-20    -0.411     0.274    -1.50    0.133
## 3 Sleep_disc20-40    -0.233     0.266    -0.877   0.381
```

For the `Sleep` variable, I’m not seeing a linear relationship. So I’m gonan leave it encoded as a categorical variable

```
coxfull<-coxph(Surv(Longevity, Censor)~ Partners+Type+Thorax+Sleep_disc, data=fruitfly_edit)
coxfull %>% tidy()
```

```
## # A tibble: 6 x 5
##   term             estimate std.error statistic  p.value
##   <chr>               <dbl>     <dbl>     <dbl>    <dbl>
## 1 Partners           0.0553    0.0304     1.82  6.86e- 2
## 2 Type               0.0336    0.0297     1.13  2.58e- 1
## 3 Thorax           -11.7       1.53      -7.68  1.56e-14
## 4 Sleep_disc>40     -0.0798    0.373     -0.214 8.31e- 1
## 5 Sleep_disc10-20   -0.151     0.279     -0.542 5.88e- 1
## 6 Sleep_disc20-40    0.208     0.270      0.771 4.41e- 1
```

```
####test sleep_disc first
coxred<-coxph(Surv(Longevity, Censor)~ Partners+Type+Thorax, data=fruitfly_edit)
G=2*(glance(coxfull)$logLik-glance(coxred)$logLik)
pchisq(G, df=3, lower.tail=FALSE)
```

```
## [1] 0.4069425
```

```
#p-value is big, so we can remove Sleep from the model

####test Type next
coxfull<-coxph(Surv(Longevity, Censor)~ Partners+Type+Thorax, data=fruitfly_edit)
coxred<-coxph(Surv(Longevity, Censor)~ Partners+Thorax, data=fruitfly_edit)
G=2*(glance(coxfull)$logLik-glance(coxred)$logLik)
pchisq(G, df=1, lower.tail=FALSE) #p-value is big so we can exclude Type from the model
```

```
## [1] 0.3461333
```

```
####test Partners next
coxfull<-coxph(Surv(Longevity, Censor)~ Partners+Thorax, data=fruitfly_edit)
coxred<-coxph(Surv(Longevity, Censor)~ Thorax, data=fruitfly_edit)
G=2*(glance(coxfull)$logLik-glance(coxred)$logLik)
pchisq(G, df=1, lower.tail=FALSE) #p-value is big so we can exclude Partners from our model from the mo
```

```
## [1] 0.1055606
```

```
####test thorax next
coxfull<-coxph(Surv(Longevity, Censor) ~ Thorax, data=fruitfly_edit)
coxred<-coxph(Surv(Longevity, Censor) ~ 1, data=fruitfly_edit)
G=2*(glance(coxfull)$logLik-glance(coxred)$logLik)
pchisq(G, df=1, lower.tail=FALSE) #p-value is tiny so we should retain Thorax in our model
```

```
## [1] 2.735121e-14
```

```
#### Final
coxfinal<-coxph(Surv(Longevity, Censor)~ Thorax, data=fruitfly_edit)
```

(c) using the final model, interpret each of the coefficients (in terms of hazard ratios). Don't forget that when a model has multiple variables, the coefficient estimate will be interpreted while keeping all other variables constant.

```
coxfinal %>% tidy()
```

```
## # A tibble: 1 x 5
##   term   estimate std.error statistic  p.value
##   <chr>     <dbl>     <dbl>     <dbl>    <dbl>
## 1 Thorax    -11.7      1.50     -7.79 6.66e-15
```

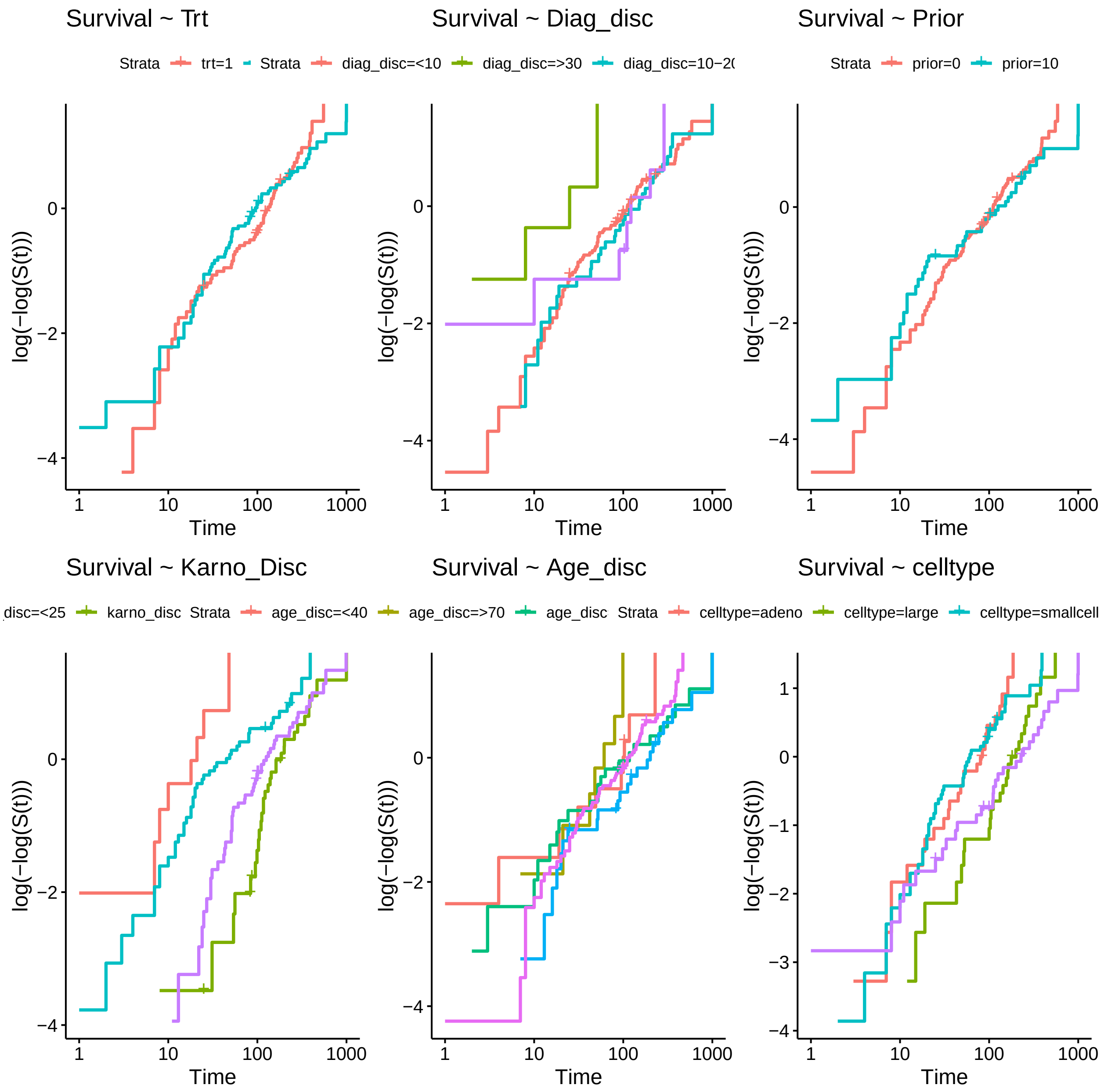
The hazard ratio associated with a 1 unit increase in Thorax length is $\exp(-11.7) = 8.3 \times 10^{-6}$. That's crazy! Looks like having a big thorax is really good for survival!

Q3. Chp 9, E12 Read text for description of data.

```
VAlung <- read_csv("https://pomona.box.com/shared/static/r6hoo1gawopkt0526xvwwze5fl3245de",
                     na="*")
```

(a) check proportional hazard assumptions for treatment KM curves. Use `fun="cloglog"` inside the `ggsurvplot`. And/or use `cox.zph`.

```
#VAlung

#########
VAlung_edit<-VAlung %>% mutate(karno_disc=ifelse(karno<25, "<25", ifelse(between(karno, 25, 50), "25-50
#########

surv1<-survival::survfit(Surv(time, status)~ trt, data=VAlung_edit)
surv2<-survival::survfit(Surv(time, status)~ karno_disc, data=VAlung_edit)
surv3<-survival::survfit(Surv(time, status)~ diag_disc, data=VAlung_edit)
surv4<-survival::survfit(Surv(time, status)~ age_disc, data=VAlung_edit)
surv5<-survival::survfit(Surv(time, status)~ prior, data=VAlung_edit)
surv6<-survival::survfit(Surv(time, status)~ celltype, data=VAlung_edit)

p1<-survminer::ggsurvplot(surv1, censor=TRUE, fun="cloglog")+ggtitle("Survival ~ Trt")
p2<-survminer::ggsurvplot(surv2, censor=TRUE, fun="cloglog")+ggtitle("Survival ~ Karno_Disc")
p3<-survminer::ggsurvplot(surv3, censor=TRUE, fun="cloglog")+ggtitle("Survival ~ Diag_disc")
p4<-survminer::ggsurvplot(surv4, censor=TRUE, fun="cloglog")+ggtitle("Survival ~ Age_disc")
p5<-survminer::ggsurvplot(surv5, censor=TRUE, fun="cloglog")+ggtitle("Survival ~ Prior")
p6<-survminer::ggsurvplot(surv6, censor=TRUE, fun="cloglog")+ggtitle("Survival ~ celltype")

arrange_ggsurvplots(list(p1, p2, p3, p4, p5, p6), ncol=3, nrow=2)
```



All the different survival curves are *roughly* parallel, indicating that proportional hazards is a reasonable assumption.

(b) use all explanatory variables and likelihood ratio test to come up with the "best" model.

Let's first test whether any of these variables should be continuous

```
coxtest1<-coxph(Surv(time, status)~karno_disc, data=VAlung_edit)
coxtest1 %>% tidy()
```

```
## # A tibble: 3 x 5
##   term              estimate std.error statistic      p.value
##   <chr>                <dbl>     <dbl>     <dbl>        <dbl>
## 1 karno_disc>75       -2.39     0.431     -5.55 0.0000000293
## 2 karno_disc25-50    -1.35     0.406     -3.31 0.000919
## 3 karno_disc50-75    -2.07     0.410     -5.06 0.000000421
```

Karno score estimates grow *roughly* linearly, so we will not include them as a continous variable.

```
coxtest2<-coxph(Surv(time, status)~diag_disc, data=VAlung_edit)
coxtest2 %>% tidy()
```

```
## # A tibble: 3 x 5
##   term           estimate std.error statistic p.value
##   <chr>             <dbl>     <dbl>     <dbl>   <dbl>
## 1 diag_disc>30     1.53       0.524      2.91  0.00360
## 2 diag_disc10-20  -0.0880     0.213     -0.412 0.680
## 3 diag_disc20-30  -0.0710     0.395     -0.180 0.857
```

The diagonal times are not linear. So I will encode it as discrete.

```
coxtest3<-coxph(Surv(time, status)~age_disc, data=VAlung_edit)
coxtest3 %>% tidy()
```

```
## # A tibble: 4 x 5
##   term          estimate std.error statistic p.value
##   <chr>            <dbl>     <dbl>     <dbl>   <dbl>
## 1 age_disc>70      0.438     0.496     0.883  0.377
## 2 age_disc40-50   -0.436     0.391    -1.12   0.265
## 3 age_disc50-60   -0.730     0.392    -1.86   0.0625
## 4 age_disc60-70   -0.273     0.342    -0.798  0.425
```

I'm going to encode age as a continous variable, since 40-50, 50-60, and 60-70 show approximately the same differences in their estimates.

```
coxfull<-coxph(Surv(time, status)~ trt+karno+diag_disc+age+prior+celltype, data=VAlung_edit)
coxfull %>% tidy()
```

```
## # A tibble: 10 x 5
##    term              estimate std.error statistic      p.value
##    <chr>                <dbl>     <dbl>     <dbl>        <dbl>
##  1 trt                0.269     0.213       1.26  0.206
##  2 karno             -0.0315    0.00561    -5.61  0.0000000201
##  3 diag_disc>30       0.648     0.579       1.12  0.262
##  4 diag_disc10-20     0.0989    0.239       0.414 0.679
##  5 diag_disc20-30    -0.0823    0.446      -0.184 0.854
##  6 age               -0.00828   0.00941    -0.880 0.379
##  7 prior              0.00288   0.0239      0.121 0.904
##  8 celltypelarge     -0.862     0.318      -2.71  0.00677
##  9 celltypesmallcell -0.381     0.276      -1.38  0.168
## 10 celltypesquamous  -1.24      0.306      -4.07  0.0000466
```

```
####test prior first
coxred<-coxph(Surv(time, status)~ trt+karno+diag_disc+age+celltype, data=VAlung_edit)
G=2*(glance(coxfull)$logLik-glance(coxred)$logLik)
pchisq(G, df=1, lower.tail=FALSE)
```

```
## [1] 0.9039754
```

```
#p-value is big (0.9), so we can remove Sleep from the model

####test diag_disc next
coxfull<-coxph(Surv(time, status)~ trt+karno+diag_disc+age+celltype, data=VAlung_edit)
coxred<-coxph(Surv(time, status)~ trt+karno+age+celltype, data=VAlung_edit)
G=2*(glance(coxfull)$logLik-glance(coxred)$logLik)
pchisq(G, df=3, lower.tail=FALSE) #p-value is big (0.69) so we can exclude Type from the model
```

```
## [1] 0.6918177
```

```
####test age next
coxfull<-coxph(Surv(time, status)~ trt+karno+age+celltype, data=VAlung_edit)
coxred<-coxph(Surv(time, status)~ trt+karno+celltype, data=VAlung_edit)
G=2*(glance(coxfull)$logLik-glance(coxred)$logLik)
pchisq(G, df=1, lower.tail=FALSE) #p-value is big (0.34) so we can exclude Partners from our model from
```

```
## [1] 0.339204
```

```
####test trt next
coxfull<-coxph(Surv(time, status)~ trt+karno+celltype, data=VAlung_edit)
coxred<-coxph(Surv(time, status)~ karno+celltype, data=VAlung_edit)
G=2*(glance(coxfull)$logLik-glance(coxred)$logLik)
pchisq(G, df=1, lower.tail=FALSE) #p-value is big (0.19) so we can exclude Partners from our model from
```

```
## [1] 0.192626
```

```
####test celltype next
coxfull<-coxph(Surv(time, status)~ karno+celltype, data=VAlung_edit)
coxred<-coxph(Surv(time, status)~ karno, data=VAlung_edit)
G=2*(glance(coxfull)$logLik-glance(coxred)$logLik)
pchisq(G, df=3, lower.tail=FALSE) #p-value is small (0.0006) so we can include celltype
```

```
## [1] 0.0006014567
```

```
####test karno next
coxfull<-coxph(Surv(time, status)~ karno+celltype, data=VAlung_edit)
coxred<-coxph(Surv(time, status)~ celltype, data=VAlung_edit)
G=2*(glance(coxfull)$logLik-glance(coxred)$logLik)
pchisq(G, df=1, lower.tail=FALSE) #p-value is small (4x10^(-9)) so we can include celltype
```

```
## [1] 4.21238e-09
```

```
#### Final
coxfinal<-coxph(Surv(time, status)~ karno+celltype, data=VAlung_edit)
```

(c) using the final model, interpret each of the coefficients (in terms of hazard ratios). Don't forget that when a model has multiple variables, the coefficient estimate will be interpreted while keeping all other variables constant.

```
coxfinal %>% tidy()
```

```
## # A tibble: 4 x 5
##   term               estimate std.error statistic       p.value
##   <chr>                 <dbl>     <dbl>     <dbl>         <dbl>
## 1 karno               -0.0311   0.00518     -6.00 0.00000000199
## 2 celltypelarge       -0.832    0.293       -2.84 0.00457
## 3 celltypesmallcell   -0.442    0.255       -1.73 0.0833
## 4 celltypesquamous    -1.16     0.293       -3.95 0.0000774
```

The Hazard ratio associated with a 1 unit increase in Karnovsky score is $\exp(-0.0311) = 0.969$, suggesting that a higher karnovsky score is good for survival. Then compared to the adeno cell type, the large cell type has an associated hazard ratio of $\exp(-0.832) = 0.435$, the small cell has an associated hazard ratio of $\exp(-0.442) = 0.643$, the squamous cell type has an associated hazard ratio of $\exp(-1.16) = 0.313$.

```
praise()
```

```
## [1] "You are wondrous!"
```