HW4

Benjamin Panny

2023-11-13

Introduction

1. (60 pts) AIDS Clinical Trials Group (actg300.csv) study. Refer to p12 of our textbook. The data come from a double-blind, placebo-controlled trial that compared the three-drug regimen of indinavir (IDV), open label zidovudine (ZDV) or stavudine (d4T), and lamivudine (3TC) with the two-drug regimen of zidovudine or stavudine and lamivudine in HIV-infected patients (Hammer et al., 1997). Patients were eligible for the trial if they had no more than 200 CD4 cells per cubic millimeter and at least three months of prior zidovudine therapy. Randomization was stratified by CD4 cell count at the time of screening. The primary outcome measure was time to AIDS defining event or death. The primary interested covariate is treatment (whether the three-drug regimen is more effective compared to the two-drug regimen). In this problem, we consider the following covariates:

Variable name Description Coding

1. tx Treatment indicator 1=treatment includes IDV, 0=treatment does not include IDV
2. sex 1=male, 2=female
3. ivdrug IV drug use history 1=never, 2=currently, 3=previously
4. karnof Karnofsky Performance:
5. Scale 100 = Normal; no complaint; no evidence of disease
6. 90 = Normal activity possible; minor signs/symptoms of disease
7. 80 = Normal activity with effort; some signs/symptoms of disease
8. 70 = Cares for self; normal activity/ active work not possible
9. CD4 Baseline CD4 count cells/milliliter
10. priorzdv Months of prior ZDV use months
11. age Age at enrollment years

```
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
```

```
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))
act <- act %>%
  select(id, time, censor, tx, sex, ivdrug, karnof, cd4, priorzdv, age) %>%
  mutate(sex = factor(if_else(sex == 1, "male", "female")),
         ivdrug = factor(case_when(
           ivdrug == 1 ~ "never",
           ivdrug == 2 ~ "currently",
           ivdrug == 3 ~ "previously",
           .default = NA
         ), levels = c("never", "currently", "previously")),
         karnof = factor(karnof,levels = c(70, 80, 90, 100)),
         )
act
```

```
## # A tibble: 1,151 × 10
##       id  time censor    tx sex    ivdrug     karnof   cd4 priorzdv   age
##    <dbl> <dbl>  <dbl> <dbl> <fct>  <fct>      <fct>  <dbl>    <dbl> <dbl>
##  1     1   189      0     0 male   never      100    169         39    34
##  2     2   287      0     0 female never      90     150.        15    34
##  3     3   242      0     1 male   never      100     23.5        9    20
##  4     4   199      0     0 male   never      90      46         53    48
##  5     5   286      0     1 male   previously 90      10         12    46
##  6     6   285      0     1 male   never      70       0         24    51
##  7     7   270      0     0 male   never      100     54.5        6    51
##  8     8   285      0     1 male   previously 80     118.        24    40
##  9     9   276      0     0 male   never      100     95          7    34
## 10    10   306      0     0 male   never      90      71          7    38
## # ℹ 1,141 more rows
```

Using the purposeful selection method for model building (taught in Chapter 5), develop your best model for evaluating the treatment effect on survival time to AIDS diagnosis or death. This process should include the following steps: • (main effect) variable selection (15 pts),

step 0: fit all univaraible Cox PH Models

```
library(survival)
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
fit.1 <- coxph(Surv(time/365.25, censor)~ tx, data=act)
fit.2 <- coxph(Surv(time/365.25, censor)~ sex, data=act)
fit.3 <- coxph(Surv(time/365.25, censor)~ ivdrug, data=act)
fit.4 <- coxph(Surv(time/365.25, censor)~ karnof, data=act)
fit.5 <- coxph(Surv(time/365.25, censor)~ cd4, data=act)
fit.6 <- coxph(Surv(time/365.25, censor)~ priorzdv, data=act)
fit.7 <- coxph(Surv(time/365.25, censor)~ age, data=act)
sum1 <- summary(fit.1)
sum2 <- summary(fit.2)
sum3 <- summary(fit.3)
sum4 <- summary(fit.4)
sum5 <- summary(fit.5)
sum6 <- summary(fit.6)
sum7 <- summary(fit.7)
rbind(sum1$coefficients, sum2$coefficients, sum3$coefficients, sum4$coefficients, sum5$coefficients, sum6$coef
ficients, sum7$coefficients) %>% kable() %>% kable_styling()
```

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
tx	-0.6844423	0.5043715	0.2149188	-3.1846556	0.0014493
sexmale	0.0791581	1.0823754	0.2811359	0.2815651	0.7782770
ivdrugcurrently	0.9876170	2.6848290	1.0059047	0.9818197	0.3261887
ivdrugpreviously	-0.4749416	0.6219214	0.3343346	-1.4205576	0.1554454
karnof80	-0.6775406	0.5078645	0.3637763	-1.8625200	0.0625298
karnof90	-1.6316053	0.1956153	0.3555343	-4.5891637	0.0000045
karnof100	-2.1099481	0.1212443	0.3987013	-5.2920521	0.0000001
cd4	-0.0161973	0.9839332	0.0025028	-6.4716925	0.0000000
priorzdv	-0.0025242	0.9974790	0.0038421	-0.6569810	0.5111931
age	0.0203405	1.0205488	0.0108383	1.8767177	0.0605568

Step 1: Using alpha = 0.2 cutoff for the univariate models, fit multivariable model

```
fit.mv1 <- coxph(Surv(time/365.25, censor)~ tx + ivdrug + karnof + cd4 + age, data=act)
sum.mv1 <- summary(fit.mv1)
sum.mv1
```

```
## Call:
## coxph(formula = Surv(time/365.25, censor) ~ tx + ivdrug + karnof +
##     cd4 + age, data = act)
##
##   n= 1151, number of events= 96
##
##                        coef exp(coef)  se(coef)      z Pr(>|z|)
## tx                -0.677589  0.507840  0.215636 -3.142 0.001676 **
## ivdrugcurrently    0.726177  2.067162  1.023934  0.709 0.478199
## ivdrugpreviously -0.619604  0.538158  0.336191 -1.843 0.065327 .
## karnof80          -0.428587  0.651429  0.367088 -1.168 0.242995
## karnof90          -1.113444  0.328426  0.367175 -3.032 0.002426 **
## karnof100         -1.548951  0.212471  0.410050 -3.777 0.000158 ***
## cd4               -0.014580  0.985525  0.002549 -5.721 1.06e-08 ***
## age                0.021187  1.021413  0.011315  1.873 0.061132 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                  exp(coef) exp(-coef) lower .95 upper .95
## tx                  0.5078     1.9691   0.33280    0.7750
## ivdrugcurrently     2.0672     0.4838   0.27784   15.3797
## ivdrugpreviously    0.5382     1.8582   0.27845    1.0401
## karnof80            0.6514     1.5351   0.31725    1.3376
## karnof90            0.3284     3.0448   0.15992    0.6745
## karnof100           0.2125     4.7065   0.09512    0.4746
## cd4                 0.9855     1.0147   0.98061    0.9905
## age                 1.0214     0.9790   0.99901    1.0443
##
## Concordance= 0.783  (se = 0.023 )
## Likelihood ratio test= 103.9  on 8 df,   p=<2e-16
## Wald test            = 85.89  on 8 df,   p=3e-15
## Score (logrank) test = 103.6  on 8 df,   p=<2e-16
```

Step 2: Use the p-values from the Wald tests of individual coefficients to identify coefficients to try deleting from the model. Confirm their removal with the partial likelihood ratio test.

```
fit.mv2 <- coxph(Surv(time/365.25, censor)~ tx + karnof + cd4, data=act)
sum.mv2 <- summary(fit.mv2)
sum.mv2
```

```
## Call:
## coxph(formula = Surv(time/365.25, censor) ~ tx + karnof + cd4,
##     data = act)
##
##   n= 1151, number of events= 96
##
##                coef exp(coef)  se(coef)      z Pr(>|z|)
## tx        -0.646524  0.523863  0.215381 -3.002  0.00268 **
## karnof80  -0.446976  0.639559  0.364834 -1.225  0.22052
## karnof90  -1.142121  0.319141  0.360471 -3.168  0.00153 **
## karnof100 -1.589237  0.204081  0.404259 -3.931 8.45e-05 ***
## cd4       -0.014031  0.986066  0.002492 -5.631 1.80e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## tx           0.5239      1.909   0.34347    0.7990
## karnof80     0.6396      1.564   0.31285    1.3075
## karnof90     0.3191      3.133   0.15745    0.6469
## karnof100    0.2041      4.900   0.09241    0.4507
## cd4          0.9861      1.014   0.98126    0.9909
##
## Concordance= 0.777  (se = 0.023 )
## Likelihood ratio test= 95.84  on 5 df,   p=<2e-16
## Wald test            = 79.82  on 5 df,   p=9e-16
## Score (logrank) test = 97.56  on 5 df,   p=<2e-16
```

```
anova(fit.mv1, fit.mv2)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(time/365.25, censor)
##  Model 1: ~ tx + ivdrug + karnof + cd4 + age
##  Model 2: ~ tx + karnof + cd4
##    loglik  Chisq Df Pr(>|Chi|)
## 1 -606.49
## 2 -610.52 8.0571  3    0.04485 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
pchisq(2*(fit.mv1$loglik[2]-fit.mv2$loglik[2]), df = 1, lower.tail=F)
```

```
## [1] 0.00453268
```

The partial likelihood ratio test indicates that fuller model outperforms the simpler model. I'll therefore remove them one by one

```
fit.mv3 <- coxph(Surv(time/365.25, censor)~ tx + ivdrug + karnof + cd4, data=act)
sum.mv3 <- summary(fit.mv3)
fit.mv4 <- coxph(Surv(time/365.25, censor)~ tx + age + karnof + cd4, data=act)
sum.mv4 <- summary(fit.mv4)
anova(fit.mv1, fit.mv3)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(time/365.25, censor)
##  Model 1: ~ tx + ivdrug + karnof + cd4 + age
##  Model 2: ~ tx + ivdrug + karnof + cd4
##    loglik  Chisq Df Pr(>|Chi|)
## 1 -606.49
## 2 -608.19 3.3929  1    0.06548 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fit.mv1, fit.mv4)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(time/365.25, censor)
##  Model 1: ~ tx + ivdrug + karnof + cd4 + age
##  Model 2: ~ tx + age + karnof + cd4
##    loglik  Chisq Df Pr(>|Chi|)
## 1 -606.49
## 2 -608.74 4.5027  2     0.1053
```

```
anova(fit.mv3, fit.mv2)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(time/365.25, censor)
##  Model 1: ~ tx + ivdrug + karnof + cd4
##  Model 2: ~ tx + karnof + cd4
##    loglik  Chisq Df Pr(>|Chi|)
## 1 -608.19
## 2 -610.52 4.6642  2    0.09709 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fit.mv4, fit.mv2)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(time/365.25, censor)
##  Model 1: ~ tx + age + karnof + cd4
##  Model 2: ~ tx + karnof + cd4
##    loglik  Chisq Df Pr(>|Chi|)
## 1 -608.74
## 2 -610.52 3.5543  1    0.05939 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It appears that removing both age and ivdrug from the full model at once results in a significantly worse log-likelihood, while removing each one at a time from the model, and then again down to the original simpler model, only results in partial likelihood ratio tests that are close to significantly worse, but the corresponding chi-square test statistics are $p > 0.05$. Therefore, at step 3, I will see if the removal of any has indications of confounding as my criteria for keeping them in the model.

Step 3. Assess whether removal of the covariate in Step 2 has produced an important change (about 20%) in the coefficients of the variables remaining in the model

```
confound_check <- as_tibble(
  rbind(
    sum.mv1$coefficients, sum.mv2$coefficients, sum.mv3$coefficients, sum.mv4$coefficients
  ) %>%
    cbind(
      coefficient = c(rownames(sum.mv1$coefficients), rownames(sum.mv2$coefficients), rownames(sum.mv3$coeffic
ients), rownames(sum.mv4$coefficients))
    ) %>%
    cbind(
      model = c(rep("full", nrow(sum.mv1$coefficients)), rep("no_age_ivdrug", nrow(sum.mv2$coefficients)), rep
("no_age", nrow(sum.mv3$coefficients)), rep("no_ivdrug", nrow(sum.mv4$coefficients)))
    )
) %>%
  mutate(coef = as.numeric(coef),
         twenty_pct_lwr = ifelse(coef <= 0, 1.2*coef, .8*coef),
         twenty_pct_upr = ifelse(coef <= 0, .8*coef, 1.2*coef),
         twenty_pct_lwr = ifelse(model == "full", twenty_pct_lwr, NA),
         twenty_pct_upr = ifelse(model == "full", twenty_pct_upr, NA)) %>%
  group_by(coefficient) %>%
  fill(twenty_pct_lwr, twenty_pct_upr) %>%
  ungroup() %>%
  mutate(within_twenty_pct_idx = if_else(between(coef, twenty_pct_lwr, twenty_pct_upr), T, F))
confound_check %>% filter(!within_twenty_pct_idx)
```

```
## # A tibble: 1 × 10
##    coef `exp(coef)` `se(coef)` z     `Pr(>|z|)` coefficient model twenty_pct_lwr
##   <dbl> <chr>       <chr>      <chr> <chr>      <chr>       <chr>           <dbl>
## 1 0.936 2.55045478… 1.0181156… 0.91… 0.3577753… ivdrugcurr… no_a…          0.581
## # ℹ 2 more variables: twenty_pct_upr <dbl>, within_twenty_pct_idx <lgl>
```

The only coefficient that moves more than 20% with the absence of another covariate compared to its value in the “full” model is the coefficient for current IV drug use in the absence of age. However, considering that none of the kept variables are significantly confounded, and neither ivdrugs nor age alone are enough to significantly worsen the log likelihood of the model, I will consider this fact irrelevant to my main effects model.

Step 4: Add to the model, one at a time, all the variables excluded from the initial multivariable model to confirm that they are neither statistically significant nor an important confounder. The model at the conclusion of this step: preliminary main effects model

I have already completed this step when removing variables one by one.

Step 5: Examine the scale of the continuous covariates, i.e., test the hypothesis that the effect of the covariate is linear in the log-hazard, and if not, what transformation should be used to make it linear. Method of fractional polynomials (Section 5.2.1)

```
library(mfp)
```

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

```
fit.J1 <- mfp(Surv(time/365.25, censor) ~ tx + karnof + fp(cd4, df = 2, select = 0.05),
              family = cox, data = act)

print(fit.J1)
```

```
## Call:
## mfp(formula = Surv(time/365.25, censor) ~ tx + karnof + fp(cd4,
##     df = 2, select = 0.05), data = act, family = cox)
##
##
## Deviance table:
##           Resid. Dev
## Null model    1316.885
## Linear model  1221.042
## Final model   1221.042
##
## Fractional polynomials:
##           df.initial select alpha df.final power1 power2
## cd4                2   0.05  0.05        1      1      .
## karnof80           1   1.00  0.05        1      1      .
## karnof90           1   1.00  0.05        1      1      .
## karnof100          1   1.00  0.05        1      1      .
## tx                 1   1.00  0.05        1      1      .
##
##
## Transformations of covariates:
##                      formula
## tx                        tx
## karnof                karnof
## cd4    I(((cd4+0.2)/100)^1)
##
##                coef exp(coef) se(coef)      z        p
## cd4.1       -1.4031    0.2458   0.2492 -5.631 1.80e-08
## karnof80.1  -0.4470    0.6396   0.3648 -1.225 2.21e-01
## karnof90.1  -1.1421    0.3191   0.3605 -3.168 1.53e-03
## karnof100.1 -1.5892    0.2041   0.4043 -3.931 8.45e-05
## tx.1        -0.6465    0.5239   0.2154 -3.002 2.68e-03
##
## Likelihood ratio test=95.84  on 5 df, p=0 n= 1151
```

```
fit.J2 <- mfp(Surv(time/365.25, censor) ~ tx + karnof + fp(cd4, df = 4, select = 0.05),
               family = cox, data = act)
print(fit.J2)
```

```
## Call:
## mfp(formula = Surv(time/365.25, censor) ~ tx + karnof + fp(cd4,
##     df = 4, select = 0.05), data = act, family = cox)
##
##
## Deviance table:
##           Resid. Dev
## Null model    1316.885
## Linear model  1221.042
## Final model   1221.042
##
## Fractional polynomials:
##           df.initial select alpha df.final power1 power2
## cd4                4   0.05  0.05        1      1      .
## karnof80           1   1.00  0.05        1      1      .
## karnof90           1   1.00  0.05        1      1      .
## karnof100          1   1.00  0.05        1      1      .
## tx                 1   1.00  0.05        1      1      .
##
##
## Transformations of covariates:
##                     formula
## tx                       tx
## karnof                karnof
## cd4    I(((cd4+0.2)/100)^1)
##
##                coef exp(coef) se(coef)      z        p
## cd4.1       -1.4031    0.2458   0.2492 -5.631 1.80e-08
## karnof80.1  -0.4470    0.6396   0.3648 -1.225 2.21e-01
## karnof90.1  -1.1421    0.3191   0.3605 -3.168 1.53e-03
## karnof100.1 -1.5892    0.2041   0.4043 -3.931 8.45e-05
## tx.1        -0.6465    0.5239   0.2154 -3.002 2.68e-03
##
## Likelihood ratio test=95.84  on 5 df, p=0 n= 1151
```

```
summary(fit.J1); summary(fit.J2)
```

```
## Call:
## coxph(formula = Surv(time/365.25, censor) ~ I(((cd4 + 0.2)/100)^1) +
##     karnof + tx, data = act)
##
##   n= 1151, number of events= 96
##
##                            coef exp(coef) se(coef)      z Pr(>|z|)
## I(((cd4 + 0.2)/100)^1) -1.4031    0.2458   0.2492 -5.631 1.80e-08 ***
## karnof80               -0.4470    0.6396   0.3648 -1.225  0.22052
## karnof90               -1.1421    0.3191   0.3605 -3.168  0.00153 **
## karnof100              -1.5892    0.2041   0.4043 -3.931 8.45e-05 ***
## tx                     -0.6465    0.5239   0.2154 -3.002  0.00268 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                        exp(coef) exp(-coef) lower .95 upper .95
## I(((cd4 + 0.2)/100)^1)    0.2458      4.068   0.15083    0.4006
## karnof80                  0.6396      1.564   0.31285    1.3075
## karnof90                  0.3191      3.133   0.15745    0.6469
## karnof100                 0.2041      4.900   0.09241    0.4507
## tx                        0.5239      1.909   0.34347    0.7990
##
## Concordance= 0.777  (se = 0.023 )
## Likelihood ratio test= 95.84  on 5 df,   p=<2e-16
## Wald test            = 79.82  on 5 df,   p=9e-16
## Score (logrank) test = 97.56  on 5 df,   p=<2e-16
```


```
## Call:
## coxph(formula = Surv(time/365.25, censor) ~ I(((cd4 + 0.2)/100)^1) +
##     karnof + tx, data = act)
##
##   n= 1151, number of events= 96
##
##                            coef exp(coef) se(coef)      z Pr(>|z|)
## I(((cd4 + 0.2)/100)^1) -1.4031    0.2458   0.2492 -5.631 1.80e-08 ***
## karnof80               -0.4470    0.6396   0.3648 -1.225  0.22052
## karnof90               -1.1421    0.3191   0.3605 -3.168  0.00153 **
## karnof100              -1.5892    0.2041   0.4043 -3.931 8.45e-05 ***
## tx                     -0.6465    0.5239   0.2154 -3.002  0.00268 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                        exp(coef) exp(-coef) lower .95 upper .95
## I(((cd4 + 0.2)/100)^1)    0.2458      4.068   0.15083    0.4006
## karnof80                  0.6396      1.564   0.31285    1.3075
## karnof90                  0.3191      3.133   0.15745    0.6469
## karnof100                 0.2041      4.900   0.09241    0.4507
## tx                        0.5239      1.909   0.34347    0.7990
##
## Concordance= 0.777  (se = 0.023 )
## Likelihood ratio test= 95.84  on 5 df,   p=<2e-16
## Wald test            = 79.82  on 5 df,   p=9e-16
## Score (logrank) test = 97.56  on 5 df,   p=<2e-16
```


Giving the multivariable fractional polynomial more degrees of freedom does not influence the final covariate transformation used for CD4.

```
summary(fit.J1)
```

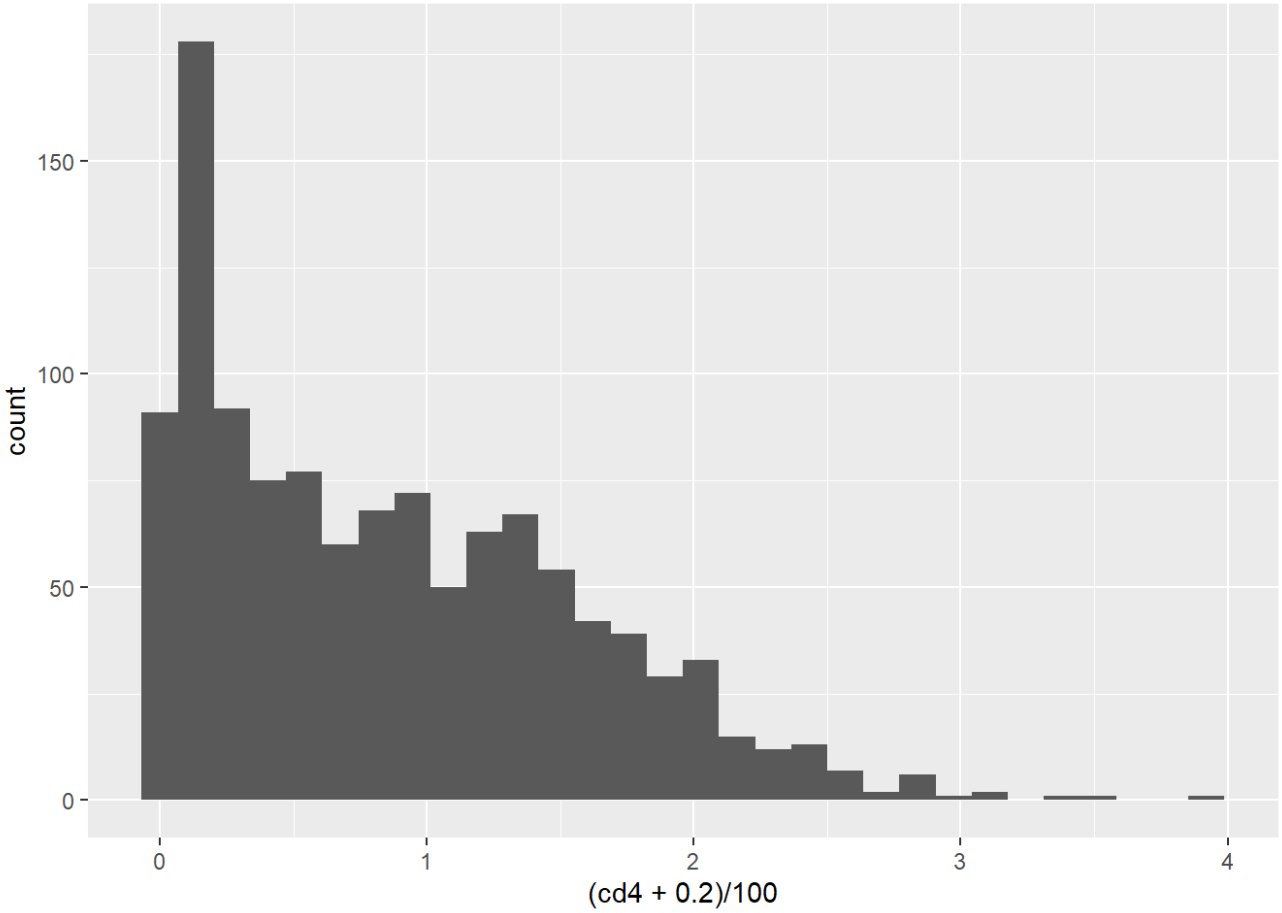
```
## Call:
## coxph(formula = Surv(time/365.25, censor) ~ I(((cd4 + 0.2)/100)^1) +
##     karnof + tx, data = act)
##
##   n= 1151, number of events= 96
##
##                            coef exp(coef) se(coef)      z Pr(>|z|)
## I(((cd4 + 0.2)/100)^1) -1.4031    0.2458   0.2492 -5.631 1.80e-08 ***
## karnof80               -0.4470    0.6396   0.3648 -1.225  0.22052
## karnof90               -1.1421    0.3191   0.3605 -3.168  0.00153 **
## karnof100              -1.5892    0.2041   0.4043 -3.931 8.45e-05 ***
## tx                     -0.6465    0.5239   0.2154 -3.002  0.00268 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                        exp(coef) exp(-coef) lower .95 upper .95
## I(((cd4 + 0.2)/100)^1)    0.2458      4.068   0.15083    0.4006
## karnof80                  0.6396      1.564   0.31285    1.3075
## karnof90                  0.3191      3.133   0.15745    0.6469
## karnof100                 0.2041      4.900   0.09241    0.4507
## tx                        0.5239      1.909   0.34347    0.7990
##
## Concordance= 0.777  (se = 0.023 )
## Likelihood ratio test= 95.84  on 5 df,   p=<2e-16
## Wald test            = 79.82  on 5 df,   p=9e-16
## Score (logrank) test = 97.56  on 5 df,   p=<2e-16
```

```
summary(fit.mv2)
```

```
## Call:
## coxph(formula = Surv(time/365.25, censor) ~ tx + karnof + cd4,
##     data = act)
##
##   n= 1151, number of events= 96
##
##                coef exp(coef)  se(coef)      z Pr(>|z|)
## tx        -0.646524  0.523863  0.215381 -3.002  0.00268 **
## karnof80  -0.446976  0.639559  0.364834 -1.225  0.22052
## karnof90  -1.142121  0.319141  0.360471 -3.168  0.00153 **
## karnof100 -1.589237  0.204081  0.404259 -3.931 8.45e-05 ***
## cd4       -0.014031  0.986066  0.002492 -5.631 1.80e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## tx           0.5239      1.909   0.34347    0.7990
## karnof80     0.6396      1.564   0.31285    1.3075
## karnof90     0.3191      3.133   0.15745    0.6469
## karnof100    0.2041      4.900   0.09241    0.4507
## cd4          0.9861      1.014   0.98126    0.9909
##
## Concordance= 0.777  (se = 0.023 )
## Likelihood ratio test= 95.84  on 5 df,   p=<2e-16
## Wald test            = 79.82  on 5 df,   p=9e-16
## Score (logrank) test = 97.56  on 5 df,   p=<2e-16
```

```
act %>% ggplot(aes(x = (cd4 + .2)/100)) + geom_histogram()
```

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



• choice of transformation forms for the continuous variables (15 pts),

The covariate transformation for CD4 appears to be that of relocating and scaling its values. This results in a scaled version of its coefficient without improving explanatory power of the model. Since this transformation doesn't improve model fit, increases model complexity, and muddies interpretation. I will use the original simplified model from which to select interactions

Step 6: Determine whether interaction terms are needed, as the product of two covariates, in the model. All interaction terms should be statistically significant at significance level of such as 1 or 5 percent.All main effects of significant interactions should remain in the model. If an insignificant interaction term is included in the model, then standard error estimates will needlessly increase, thus unnecessarily widening confidence intervals of HRs

```
fit.int1 <- coxph(formula = Surv(time/365.25, censor) ~ tx*karnof + tx*cd4 + karnof*cd4,
    data = act)
summary(fit.int1)
```

```
## Call:
## coxph(formula = Surv(time/365.25, censor) ~ tx * karnof + tx *
##     cd4 + karnof * cd4, data = act)
##
##   n= 1151, number of events= 96
##
##                     coef  exp(coef)   se(coef)      z Pr(>|z|)
## tx             0.2717982  1.3123221  0.6674916  0.407  0.68387
## karnof80       0.0148661  1.0149772  0.5409950  0.027  0.97808
## karnof90      -0.6796178  0.5068107  0.5397517 -1.259  0.20798
## karnof100     -1.6331998  0.1953036  0.6245925 -2.615  0.00893 **
## cd4           -0.0157566  0.9843668  0.0120279 -1.310  0.19020
## tx:karnof80   -1.1360936  0.3210708  0.7704598 -1.475  0.14033
## tx:karnof90   -1.0381620  0.3541049  0.7481585 -1.388  0.16525
## tx:karnof100  -0.7382328  0.4779578  0.8623454 -0.856  0.39196
## tx:cd4        -0.0005310  0.9994691  0.0053382 -0.099  0.92076
## karnof80:cd4  -0.0006373  0.9993629  0.0129450 -0.049  0.96073
## karnof90:cd4  -0.0005787  0.9994215  0.0123244 -0.047  0.96255
## karnof100:cd4  0.0075965  1.0076254  0.0124674  0.609  0.54232
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##               exp(coef) exp(-coef) lower .95 upper .95
## tx               1.3123     0.7620   0.35471    4.8552
## karnof80         1.0150     0.9852   0.35153    2.9306
## karnof90         0.5068     1.9731   0.17596    1.4598
## karnof100        0.1953     5.1202   0.05742    0.6643
## cd4              0.9844     1.0159   0.96143    1.0078
## tx:karnof80      0.3211     3.1146   0.07092    1.4535
## tx:karnof90      0.3541     2.8240   0.08171    1.5345
## tx:karnof100     0.4780     2.0922   0.08818    2.5907
## tx:cd4           0.9995     1.0005   0.98907    1.0100
## karnof80:cd4     0.9994     1.0006   0.97433    1.0250
## karnof90:cd4     0.9994     1.0006   0.97557    1.0239
## karnof100:cd4    1.0076     0.9924   0.98330    1.0326
##
## Concordance= 0.781  (se = 0.022 )
## Likelihood ratio test= 100.5  on 12 df,   p=4e-16
## Wald test            = 85.98  on 12 df,   p=3e-13
## Score (logrank) test = 123.7  on 12 df,   p=<2e-16
```

```
fit.int2 <- coxph(formula = Surv(time/365.25, censor) ~ tx*karnof + tx*cd4,
    data = act)
summary(fit.int2)
```

```
## Call:
## coxph(formula = Surv(time/365.25, censor) ~ tx * karnof + tx *
##     cd4, data = act)
##
##   n= 1151, number of events= 96
##
##                     coef  exp(coef)   se(coef)      z Pr(>|z|)
## tx            0.2401795  1.2714774  0.6477206  0.371   0.7108
## karnof80     -0.0249989  0.9753110  0.4975713 -0.050   0.9599
## karnof90     -0.7493964  0.4726518  0.4986150 -1.503   0.1329
## karnof100    -1.3028101  0.2717670  0.5476260 -2.379   0.0174 *
## cd4          -0.0141266  0.9859727  0.0031483 -4.487 7.22e-06 ***
## tx:karnof80  -1.1161225  0.3275474  0.7479336 -1.492   0.1356
## tx:karnof90  -1.0016371  0.3672777  0.7260243 -1.380   0.1677
## tx:karnof100 -0.6873209  0.5029216  0.8198179 -0.838   0.4018
## tx:cd4       -0.0003662  0.9996339  0.0052401 -0.070   0.9443
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##              exp(coef) exp(-coef) lower .95 upper .95
## tx              1.2715     0.7865   0.35725    4.5253
## karnof80        0.9753     1.0253   0.36780    2.5863
## karnof90        0.4727     2.1157   0.17788    1.2559
## karnof100       0.2718     3.6796   0.09291    0.7949
## cd4             0.9860     1.0142   0.97991    0.9921
## tx:karnof80     0.3275     3.0530   0.07562    1.4188
## tx:karnof90     0.3673     2.7227   0.08851    1.5240
## tx:karnof100    0.5029     1.9884   0.10085    2.5080
## tx:cd4          0.9996     1.0004   0.98942    1.0100
##
## Concordance= 0.778  (se = 0.023 )
## Likelihood ratio test= 98.36  on 9 df,   p=<2e-16
## Wald test            = 80.76  on 9 df,   p=1e-13
## Score (logrank) test = 108.4  on 9 df,   p=<2e-16
```

```
fit.int3 <- coxph(formula = Surv(time/365.25, censor) ~ tx*karnof + cd4,
    data = act)
summary(fit.int3)
```

```
## Call:
## coxph(formula = Surv(time/365.25, censor) ~ tx * karnof + cd4,
##     data = act)
##
##   n= 1151, number of events= 96
##
##                    coef exp(coef)  se(coef)      z Pr(>|z|)
## tx            0.230910  1.259746  0.634053  0.364   0.7157
## karnof80     -0.023103  0.977162  0.496828 -0.047   0.9629
## karnof90     -0.745489  0.474502  0.495442 -1.505   0.1324
## karnof100    -1.298844  0.272847  0.544650 -2.385   0.0171 *
## cd4          -0.014260  0.985841  0.002517 -5.665 1.47e-08 ***
## tx:karnof80  -1.118292  0.326838  0.747280 -1.496   0.1345
## tx:karnof90  -1.007472  0.365141  0.721195 -1.397   0.1624
## tx:karnof100 -0.695446  0.498852  0.811561 -0.857   0.3915
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##              exp(coef) exp(-coef) lower .95 upper .95
## tx              1.2597     0.7938   0.36356    4.3651
## karnof80        0.9772     1.0234   0.36903    2.5874
## karnof90        0.4745     2.1075   0.17969    1.2530
## karnof100       0.2728     3.6651   0.09382    0.7935
## cd4             0.9858     1.0144   0.98099    0.9907
## tx:karnof80     0.3268     3.0596   0.07555    1.4139
## tx:karnof90     0.3651     2.7387   0.08883    1.5009
## tx:karnof100    0.4989     2.0046   0.10166    2.4478
##
## Concordance= 0.778  (se = 0.023 )
## Likelihood ratio test= 98.35  on 8 df,   p=<2e-16
## Wald test            = 81.01  on 8 df,   p=3e-14
## Score (logrank) test = 105.1  on 8 df,   p=<2e-16
```

• selection of interactions (15 pts).

None of the possible interactions are statistically significant.

Therefore, my final model is:

$$h(t|X) = h_0(t)\exp(\beta_1 \cdot tx + \beta_2 \cdot karnof + \beta_3 \cdot cd4)$$

• Finally, interpret your findings regarding the treatment effects from your final model (15 pts, need to provide statistics such as hazard ratio estimates, 95% CIs, p-values, etc., to support your explanation).

```
sum.mv2
```

```
## Call:
## coxph(formula = Surv(time/365.25, censor) ~ tx + karnof + cd4,
##     data = act)
##
##   n= 1151, number of events= 96
##
##                coef exp(coef)  se(coef)      z Pr(>|z|)
## tx        -0.646524  0.523863  0.215381 -3.002  0.00268 **
## karnof80  -0.446976  0.639559  0.364834 -1.225  0.22052
## karnof90  -1.142121  0.319141  0.360471 -3.168  0.00153 **
## karnof100 -1.589237  0.204081  0.404259 -3.931 8.45e-05 ***
## cd4       -0.014031  0.986066  0.002492 -5.631 1.80e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## tx           0.5239      1.909   0.34347    0.7990
## karnof80     0.6396      1.564   0.31285    1.3075
## karnof90     0.3191      3.133   0.15745    0.6469
## karnof100    0.2041      4.900   0.09241    0.4507
## cd4          0.9861      1.014   0.98126    0.9909
##
## Concordance= 0.777  (se = 0.023 )
## Likelihood ratio test= 95.84  on 5 df,   p=<2e-16
## Wald test            = 79.82  on 5 df,   p=9e-16
## Score (logrank) test = 97.56  on 5 df,   p=<2e-16
```

The effect of treatment has a hazard ratio of 0.52 with no-treatment with 95% CI [.343, .8], indicating treatment lowers the hazard rate compared to no-treatment by 47.6% (given by exp(coef) for treatment) with the equivalent 95% CI [20%, 65.7%]. This effect is statistically significant, indicated by a Wald test statistic significantly unlikely to come from the null Z-distribution (under which the assumption is that the hazard rate is equivalent between two and three-drug regimens), as well as the fact that the 95% CI, which should capture the true hazard ratio 95% of the time if we follow the same procedures and our assumptions are correct, which does not contain zero on this occassion.

Please include your key program codes and supporting outputs. You are welcome to use plots to help your model development as well as your model interpretation.

2. (40 pts) For the final model you chose in problem 1, assess the model fit, which should include the following steps:

$$h(t|X) = h_0(t)\exp(\beta_1 \cdot tx + \beta_2 \cdot karnof + \beta_3 \cdot cd4)$$

• evaluation of the proportional hazards assumption for each main effect (10 pts),

This can be accomplished via categorical quantiles indicators for the continuous variables and plots of the categorical variables.

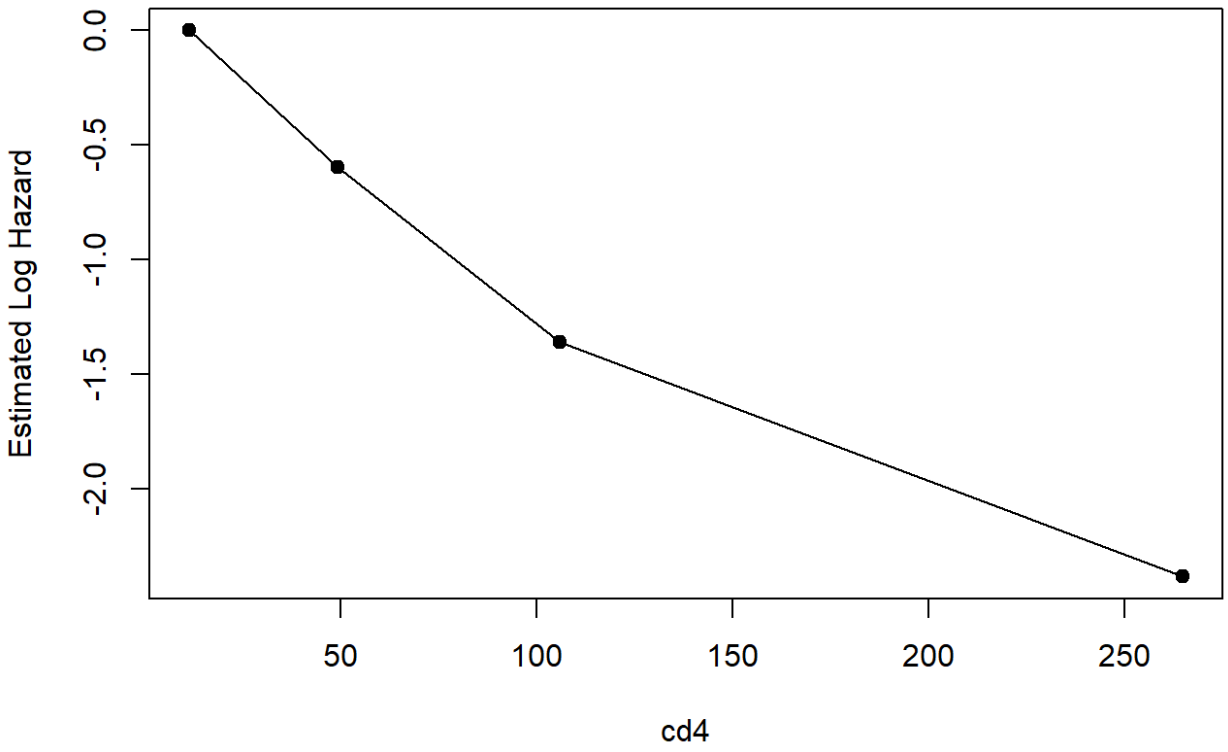
```
# compute quartile and each mid-point
cd4.q <- quantile(act$cd4)
cd4.q.diff <- cd4.q[-1] - cd4.q[-5]
cd4.q.mid <- cd4.q[-5] + 0.5*cd4.q.diff
cd4.q.mid <- c(cd4.q.mid[1],cd4.q.mid[-1]+0.5)
act$cd4.c<-cut(act$cd4,cd4.q)

fit.cd4.c <- coxph(Surv(time/365.25, censor)~ cd4.c + tx + karnof, data = act)
fit.cd4.c
```

```
## Call:
## coxph(formula = Surv(time/365.25, censor) ~ cd4.c + tx + karnof,
##     data = act)
##
##                      coef exp(coef) se(coef)      z        p
## cd4.c(23,74.5]  -0.59365   0.55231  0.23997 -2.474 0.013366
## cd4.c(74.5,136] -1.35702   0.25743  0.33898 -4.003 6.25e-05
## cd4.c(136,392]  -2.38061   0.09249  0.52240 -4.557 5.19e-06
## tx              -0.61507   0.54061  0.21664 -2.839 0.004524
## karnof80        -0.55540   0.57384  0.36460 -1.523 0.127679
## karnof90        -1.28619   0.27632  0.36318 -3.542 0.000398
## karnof100       -1.69067   0.18440  0.40561 -4.168 3.07e-05
##
## Likelihood ratio test=96.56  on 7 df, p=< 2.2e-16
## n= 1131, number of events= 94
##    (20 observations deleted due to missingness)
```

```
cd4.c.beta <- fit.cd4.c$coeff[1:3]

# Plot to examine the linearity assumption
plot(x = cd4.q.mid, y=c(0,cd4.c.beta), xlab="cd4", ylab="Estimated Log Hazard",pch=19)
lines(x = cd4.q.mid, y=c(0,cd4.c.beta),lty=1)
```



It seems as though not including the cd4 variable transformation may have violated the linearity assumption, so I will try testing the assumption with the transformation

I do not test the linearity assumption for the categorical variables because they only have 1 value, which will always be linear with 0.

• examination of the functional (transformation) form of each continuous variable being included in the model (10 pts),

I will try testing the assumption with the transformation given previously

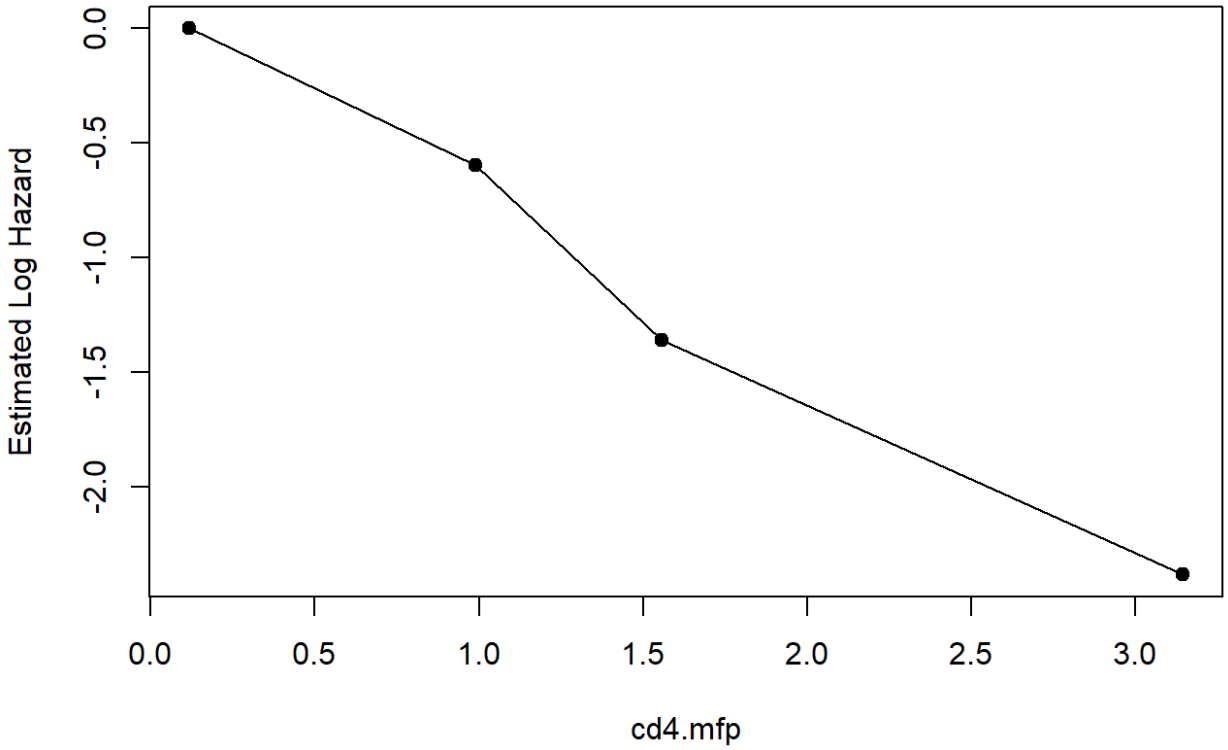
```
act$cd4.mfp <- (act$cd4 + 0.2)/100
# compute quartile and each mid-point
cd4.mfp.q <- quantile(act$cd4.mfp)
cd4.mfp.q.diff <- cd4.mfp.q[-1] - cd4.mfp.q[-5]
cd4.mfp.q.mid <- cd4.mfp.q[-5] + 0.5*cd4.mfp.q.diff
cd4.mfp.q.mid <- c(cd4.mfp.q.mid[1],cd4.mfp.q.mid[-1]+0.5)
act$cd4.mfp.c<-cut(act$cd4.mfp,cd4.mfp.q)

fit.cd4.mfp.c <- coxph(Surv(time/365.25, censor)~ cd4.mfp.c + tx + karnof, data = act)
fit.cd4.mfp.c
```

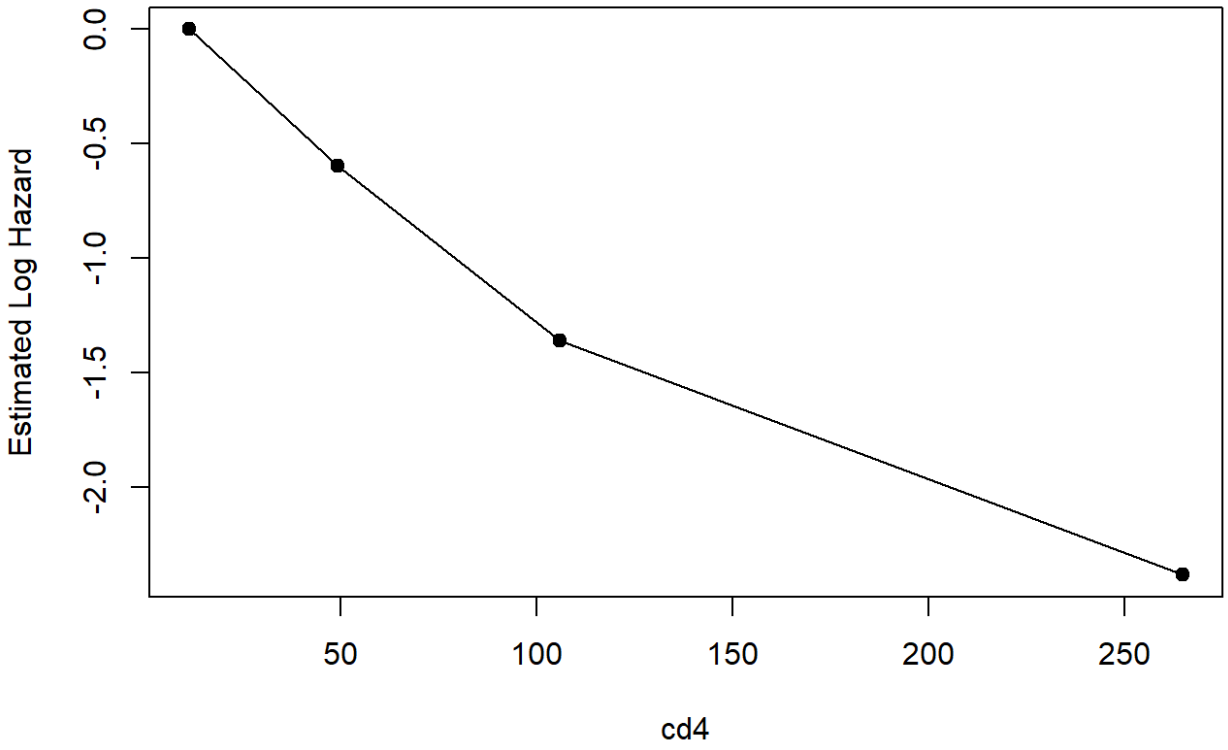
```
## Call:
## coxph(formula = Surv(time/365.25, censor) ~ cd4.mfp.c + tx +
##     karnof, data = act)
##
##                             coef exp(coef) se(coef)      z        p
## cd4.mfp.c(0.232,0.747] -0.59365   0.55231  0.23997 -2.474 0.013366
## cd4.mfp.c(0.747,1.37]  -1.35702   0.25743  0.33898 -4.003 6.25e-05
## cd4.mfp.c(1.37,3.92]   -2.38061   0.09249  0.52240 -4.557 5.19e-06
## tx                     -0.61507   0.54061  0.21664 -2.839 0.004524
## karnof80               -0.55540   0.57384  0.36460 -1.523 0.127679
## karnof90               -1.28619   0.27632  0.36318 -3.542 0.000398
## karnof100              -1.69067   0.18440  0.40561 -4.168 3.07e-05
##
## Likelihood ratio test=96.56  on 7 df, p=< 2.2e-16
## n= 1131, number of events= 94
##    (20 observations deleted due to missingness)
```

```
cd4.mfp.c.beta <- fit.cd4.mfp.c$coeff[1:3]

# Plot to examine the linearity assumption
plot(x = cd4.mfp.q.mid, y=c(0,cd4.mfp.c.beta), xlab="cd4.mfp", ylab="Estimated Log Hazard",pch=19)
lines(x = cd4.mfp.q.mid, y=c(0,cd4.mfp.c.beta),lty=1)
```




```
plot(x = cd4.q.mid, y=c(0,cd4.c.beta), xlab="cd4", ylab="Estimated Log Hazard",pch=19)
lines(x = cd4.q.mid, y=c(0,cd4.c.beta),lty=1)
```



The transformed cd4 covariate appears to be slightly more linear than the untransformed form. As I mentioned before, the cd4 transform involves a small location shift and scaling by .01.

• identification of influential subjects, if any (10 pts), and

```
## Scaled score residual -- use type="dfbeta"
score.r<-resid(fit.J1,type="dfbeta")
par(mfrow=c(3,2))
dev.off()
```

```
## null device
##           1
```

```
influentials <- tibble(cd4.mfp = act$cd4.mfp,
                       karnof = act$karnof,
                       tx = act$tx,
                       id = act$id,
                       cd4.inf = score.r[,1],
                       karnof80.inf = score.r[,2],
                       karnof90.inf = score.r[,3],
                       karnof100.inf = score.r[,4],
                       tx.inf = score.r[,5])

influentials %>%
  ggplot(aes(x = cd4.mfp, y = cd4.inf, label = id)) +
  geom_point() +
  geom_label() +
  labs(x = 'cd4', y = 'cd4 influence')

influentials %>%
  ggplot(aes(x = karnof, y = karnof80.inf, label = id)) +
  geom_point() +
  geom_label() +
  labs(x = 'karnof', y = 'karnof80 influence')

influentials %>%
  ggplot(aes(x = karnof, y = karnof90.inf, label = id)) +
  geom_point() +
  geom_label() +
  labs(x = 'karnof', y = 'karnof90 influence')

influentials %>%
  ggplot(aes(x = karnof, y = karnof100.inf, label = id)) +
  geom_point() +
  geom_label() +
  labs(x = 'karnof', y = 'karnof100 influence')

influentials %>%
  ggplot(aes(x = tx, y = tx.inf, label = id)) +
  geom_point() +
  geom_label() +
  labs(x = 'tx', y = 'tx influence')
```

853 and 967 appear highly influential with regards to the cd4 coefficient. 929 and other subjects appear influential on karnofsky scores. None of the subjects appear highly influential over the tx coefficient.

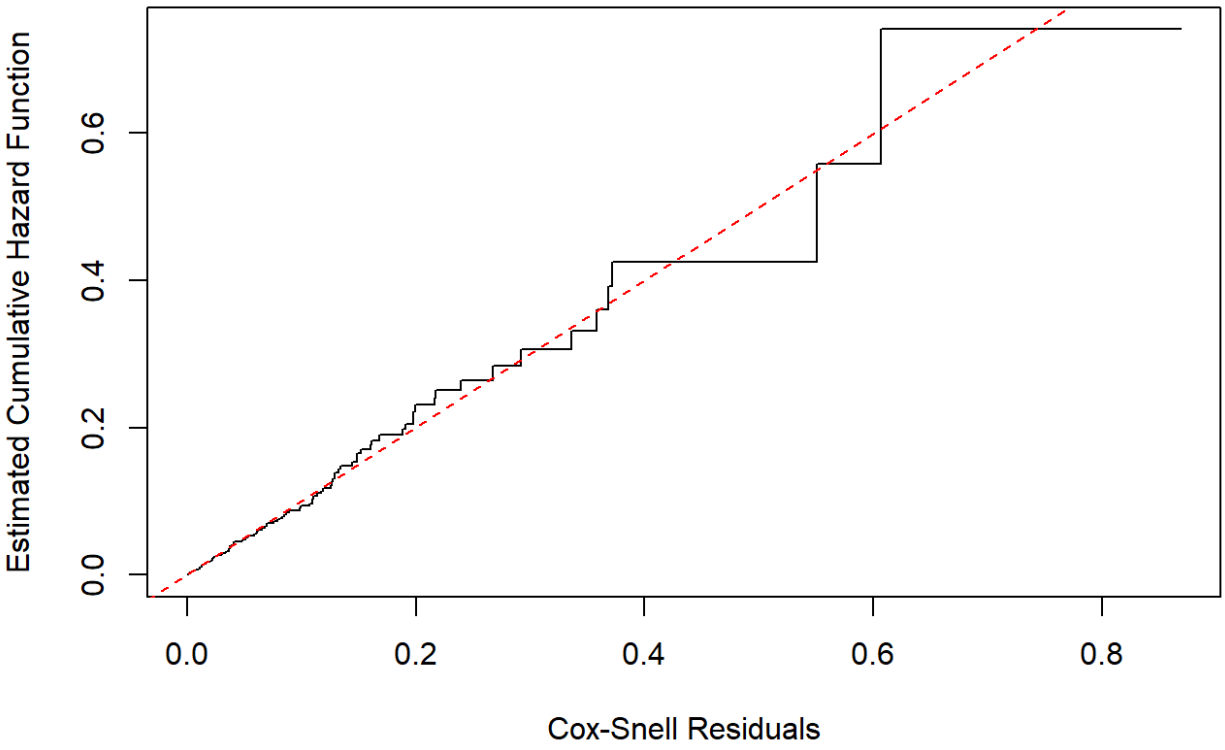
• overall model fit assessment (10pts).

As discussed for the previous question, the final model had equivalent performance to the other models including the “full” set of covariates, with a caveat. It’s performance was statistically equivalent to models that add age and ivdrug indepdendently, but statistically significantly worse compared to the “full” model. Taking these two facts together. I consider the statistically significant difference betwen the ‘full’ model and my final model to be a consequence of the fact that likelihoods always improve when more and more covariates are added to the model as opposed to the ‘full’ model having more explanatory power as a virtue of true relationships. Model fit can also be assessed with Cox-Snell residuals.

```
coxsnell.r <- act$censor-resid(fit.J1,type="martingale")

fitres <- survfit(Surv(coxsnell.r, act$censor)~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 estimate baseline hazard cumulative hazard function are close to their true values.