BIS 630 Homework 5

Joanna Chen

3/25/2020

```
setwd("~/Downloads/HW 5 survival")
library(readr)
library(data.table)
library(survival)
df <- read_csv("HW05_recidivism.csv")
```

```
## Parsed with column specification:
## cols(
##   ID = col_double(),
##   week = col_double(),
##   arrest = col_double(),
##   fin = col_double(),
##   age = col_double(),
##   race = col_double(),
##   wexp = col_double(),
##   mar = col_double(),
##   paro = col_double(),
##   prio = col_double(),
##   educ = col_double()
## )
```

1. [40 points] Let us begin by assuming a semi-final Cox model that contains FIN, AGE, PRIO, and EDUC. Represent EDUC as a categorical variable in the model (reference category = 3). In this question, you are asked to carry out Step 6 of the model selection process for AGE and PRIO and Step 7 for EDUC (see Lesson 5).

```
df$educcat = factor(df$educ, levels = c(3,4,5,6), labels = c("9th grade or less", "10th to 11th grade",
```

a. Begin by examining the scale of continuous AGE in a model that also contains FIN, PRIO, and EDUC. In a DATA step, create a categorical variable representing the four quartiles of AGE.

```
summary(df$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.00   20.00   23.00   24.62   27.00   44.00
```

```
setDT(df)
df$agecat4 = df[,ifelse(age>=17 & age<=20,1,ifelse(age>20 & age<=23,2,ifelse(age>23 & age<=27,3,ifelse(
df$agecat4 = factor(df$agecat4)
```

Include this categorical version of AGE in the model containing FIN, PRIO, and EDUC.

```
cox1 <- coxph(Surv(week, arrest) ~ fin + prio + relevel(educcat,"9th grade or less") + agecat4,data=df,
summary(cox1)
```

```
## Call:
```

```
## coxph(formula = Surv(week, arrest) ~ fin + prio + relevel(educcat,
##     "9th grade or less") + agecat4, data = df, ties = "breslow")
##
##   n= 432, number of events= 114
##
##                                                              coef exp(coef)
## fin                                                      -0.32409   0.72318
## prio                                                      0.08334   1.08691
## relevel(educcat, "9th grade or less")10th to 11th grade -0.16369   0.84901
## relevel(educcat, "9th grade or less")12th grade         -0.78636   0.45550
## relevel(educcat, "9th grade or less")Some college       -1.20838   0.29868
## agecat42                                                 -0.53694   0.58454
## agecat43                                                 -0.68152   0.50585
## agecat44                                                 -0.93294   0.39340
##                                                          se(coef)      z
## fin                                                       0.19103 -1.697
## prio                                                      0.02874  2.900
## relevel(educcat, "9th grade or less")10th to 11th grade  0.22815 -0.717
## relevel(educcat, "9th grade or less")12th grade          0.46540 -1.690
## relevel(educcat, "9th grade or less")Some college        1.01025 -1.196
## agecat42                                                  0.24009 -2.236
## agecat43                                                  0.26506 -2.571
## agecat44                                                  0.28831 -3.236
##                                                          Pr(>|z|)
## fin                                                       0.08978 .
## prio                                                      0.00374 **
## relevel(educcat, "9th grade or less")10th to 11th grade  0.47309
## relevel(educcat, "9th grade or less")12th grade          0.09109 .
## relevel(educcat, "9th grade or less")Some college        0.23165
## agecat42                                                  0.02532 *
## agecat43                                                  0.01013 *
## agecat44                                                  0.00121 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                                          exp(coef) exp(-coef)
## fin                                                         0.7232      1.383
## prio                                                        1.0869      0.920
## relevel(educcat, "9th grade or less")10th to 11th grade    0.8490      1.178
## relevel(educcat, "9th grade or less")12th grade            0.4555      2.195
## relevel(educcat, "9th grade or less")Some college          0.2987      3.348
## agecat42                                                    0.5845      1.711
## agecat43                                                    0.5058      1.977
## agecat44                                                    0.3934      2.542
##                                                          lower .95 upper .95
## fin                                                        0.49733    1.0516
## prio                                                       1.02738    1.1499
## relevel(educcat, "9th grade or less")10th to 11th grade   0.54289    1.3277
## relevel(educcat, "9th grade or less")12th grade           0.18295    1.1341
## relevel(educcat, "9th grade or less")Some college         0.04124    2.1634
## agecat42                                                   0.36513    0.9358
## agecat43                                                   0.30089    0.8504
## agecat44                                                   0.22357    0.6922
##
```

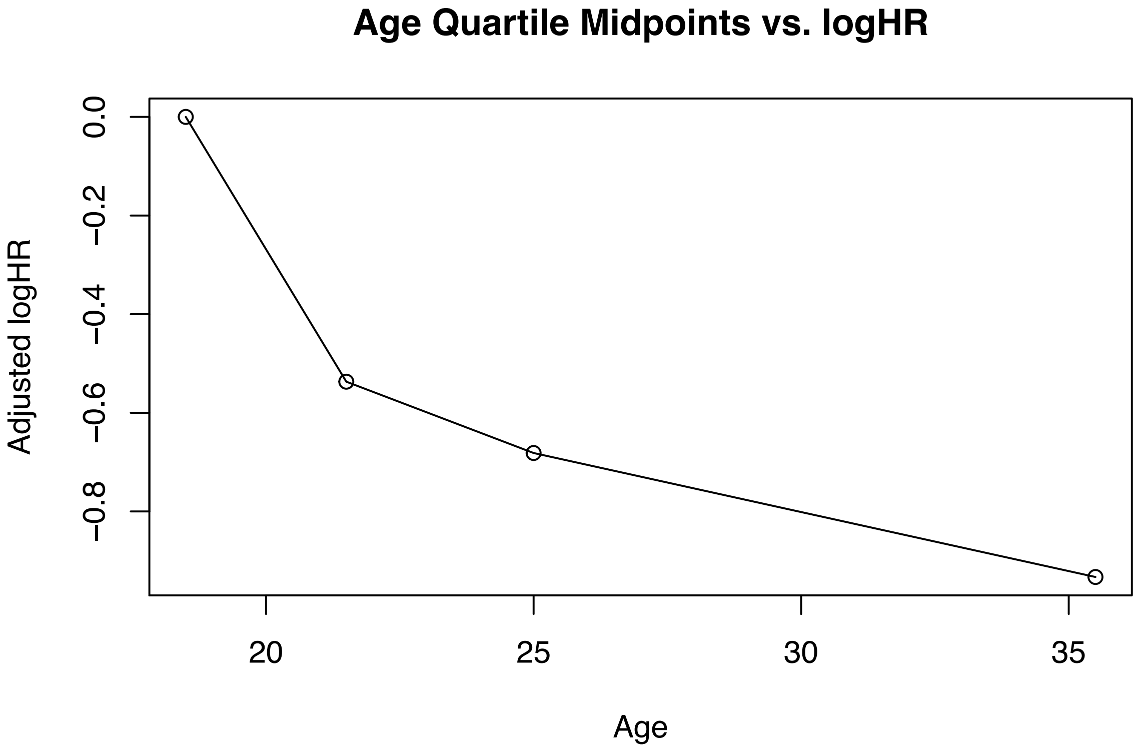
```
## Concordance= 0.659  (se = 0.026 )
## Likelihood ratio test= 37.44  on 8 df,   p=1e-05
## Wald test            = 36.65  on 8 df,   p=1e-05
## Score (logrank) test = 39.21  on 8 df,   p=4e-06
```

Plot the log hazard ratio of each category versus the midpoints of the age interval.

```
c((17+20)/2, (20+23)/2,(23+27)/2,(27+44)/2)
```

```
## [1] 18.5 21.5 25.0 35.5
```

```
x = c(18.5, 21.5, 25.0, 35.5)
y = c(0,-0.53694, -0.68152, -0.93294)
#plot(x,y,xlab="Age",ylab="Adjusted logHR", main="Age Quartile Midpoints vs. logHR",type = "l")
plot(x,y,xlab="Age",ylab="Adjusted logHR", main="Age Quartile Midpoints vs. logHR")
lines(x,y,xlim=range(x),ylim=range(y))
```



Does this plot appear fairly linear (or at least monotonically increasing/decreasing as the covariate increases)?

Yes, the plot appear at least monotonically decreasing as the age quartile midpoints increase in the log hazards. If there's more points, it's likely to be linear.

Based on this plot, is the assumption of linearity of AGE justified?

Yes. the plot is monotone decreasing. Although it may look not strictly, but note that there are only four points. I feel comfortable to include it as a linear continuous variable in my model.

b. Repeat (a) for PRIO (in a model that also contains FIN, AGE (continuous) and EDUC).

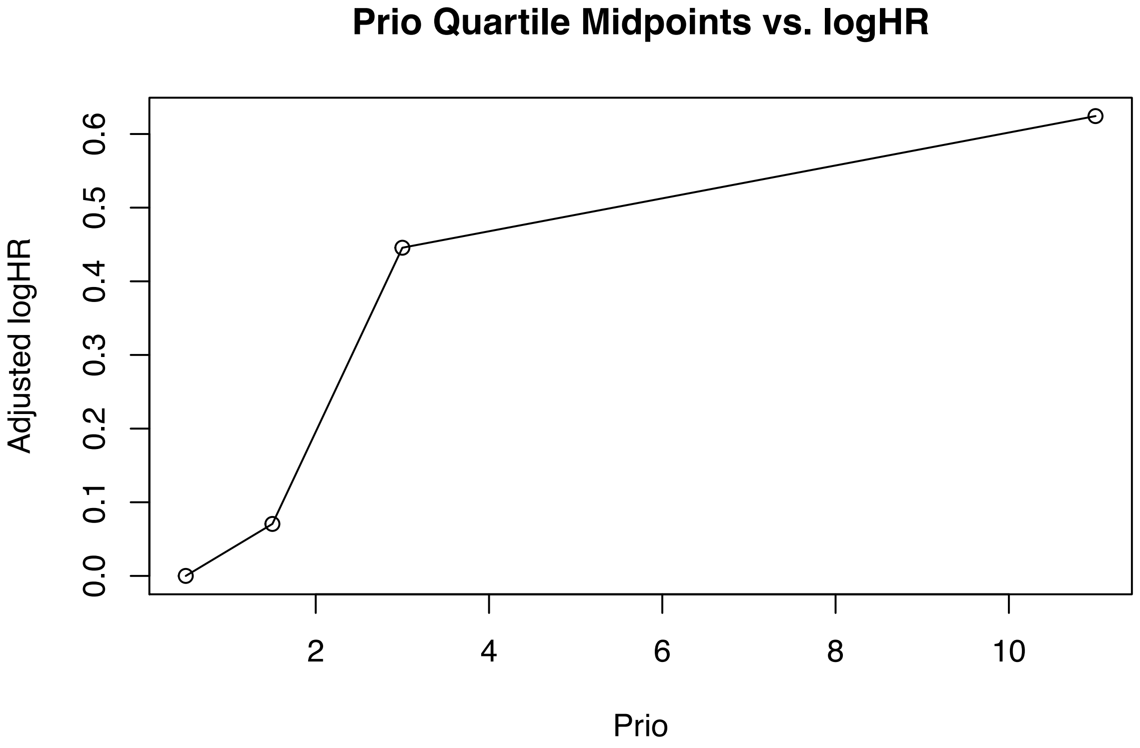
```
summary(df$prio)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.000   1.000   2.000   2.984   4.000  18.000
```

```
df$priocat4 = df[,ifelse(prio>=0 & prio<=1,1,ifelse(prio>1 & prio<=2,2,ifelse(prio>2 & prio<=4,3,ifelse
df$priocat4 = factor(df$priocat4)

cox2 <- coxph(Surv(week, arrest) ~ fin + priocat4 + relevel(educcat,"9th grade or less") + age,data=df,
summary(cox2)
```

```
## Call:
## coxph(formula = Surv(week, arrest) ~ fin + priocat4 + relevel(educcat, 
##     "9th grade or less") + age, data = df, ties = "breslow")
## 
##   n= 432, number of events= 114 
## 
##                                                            coef exp(coef)
## fin                                                    -0.30526   0.73693
## priocat42                                               0.07056   1.07311
## priocat43                                               0.44557   1.56138
## priocat44                                               0.62430   1.86694
## relevel(educcat, "9th grade or less")10th to 11th grade -0.19422   0.82348
## relevel(educcat, "9th grade or less")12th grade        -0.75404   0.47046
## relevel(educcat, "9th grade or less")Some college      -1.20835   0.29869
## age                                                    -0.05994   0.94182
##                                                        se(coef)      z
## fin                                                     0.19088 -1.599
## priocat42                                               0.28492  0.248
## priocat43                                               0.25901  1.720
## priocat44                                               0.26404  2.364
## relevel(educcat, "9th grade or less")10th to 11th grade  0.22929 -0.847
## relevel(educcat, "9th grade or less")12th grade         0.46576 -1.619
## relevel(educcat, "9th grade or less")Some college       1.01045 -1.196
## age                                                     0.02033 -2.949
##                                                        Pr(>|z|)   
## fin                                                     0.10977   
## priocat42                                               0.80442   
## priocat43                                               0.08538 . 
## priocat44                                               0.01806 * 
## relevel(educcat, "9th grade or less")10th to 11th grade  0.39696   
## relevel(educcat, "9th grade or less")12th grade         0.10546   
## relevel(educcat, "9th grade or less")Some college       0.23175   
## age                                                     0.00319 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
##                                                        exp(coef) exp(-coef)
## fin                                                       0.7369     1.3570
## priocat42                                                 1.0731     0.9319
## priocat43                                                 1.5614     0.6405
## priocat44                                                 1.8669     0.5356
## relevel(educcat, "9th grade or less")10th to 11th grade    0.8235     1.2144
## relevel(educcat, "9th grade or less")12th grade           0.4705     2.1256
## relevel(educcat, "9th grade or less")Some college          0.2987     3.3480
## age                                                        0.9418     1.0618
##                                                         lower .95 upper .95
## fin                                                       0.50693    1.0713
## priocat42                                                 0.61392    1.8757
## priocat43                                                 0.93980    2.5941
## priocat44                                                 1.11270    3.1324
## relevel(educcat, "9th grade or less")10th to 11th grade   0.52539    1.2907
## relevel(educcat, "9th grade or less")12th grade           0.18883    1.1721
## relevel(educcat, "9th grade or less")Some college         0.04122    2.1643
## age                                                       0.90503    0.9801
##
## Concordance= 0.661  (se = 0.026 )
## Likelihood ratio test= 33.02  on 8 df,   p=6e-05
## Wald test            = 28.94  on 8 df,   p=3e-04
## Score (logrank) test = 30.32  on 8 df,   p=2e-04
```

```
x = c((0+1)/2, (1+2)/2,(2+4)/2,(4+18)/2)
y = c(0,0.07056, 0.44557, 0.62430)
plot(x,y,xlab="Prio",ylab="Adjusted logHR", main="Prio Quartile Midpoints vs. logHR")
lines(x,y,xlim=range(x),ylim=range(y))
```



The plot appear at least monotonically increasing as the prio quartile midpoints increase in the log hazards. Although it may look not strictly, but note that there are only four points. Therefore, I feel comfortable to include it as a linear continuous variable in my model.

c. Notice that EDUC is an ordinal variable. Rather than treat EDUC as a nominal categorical variable, would it be appropriate to treat EDUC as a quantitative/numerical variable in a model also containing FIN, AGE, and PRIO? Follow Step 7 of the model-building process to

answer this question.

To answer this, we will test whether there is a departure from linear trend. We can use a likelihood ratio test to compare the fit of the linear model (reduced, cox4) to the fit of the nominal model (full, cox3). The slides lecture 5 page 23 shows that the linear model is nested in the categorical model.

```
cox3 <- coxph(Surv(week, arrest) ~ fin + age + prio + relevel(educcat,"9th grade or less"),data=df, tie
cox4 <- coxph(Surv(week, arrest) ~ fin + age + prio + educ,data=df, ties="breslow")
summary(cox3)
```

```
## Call:
## coxph(formula = Surv(week, arrest) ~ fin + age + prio + relevel(educcat, 
##     "9th grade or less"), data = df, ties = "breslow")
## 
##   n= 432, number of events= 114 
## 
##                                                             coef exp(coef)
## fin                                                     -0.33303   0.71675
## age                                                     -0.06094   0.94088
## prio                                                     0.08319   1.08675
## relevel(educcat, "9th grade or less")10th to 11th grade -0.21565   0.80602
## relevel(educcat, "9th grade or less")12th grade         -0.75775   0.46872
## relevel(educcat, "9th grade or less")Some college       -1.23629   0.29046
##                                                         se(coef)      z
## fin                                                      0.19072 -1.746
## age                                                      0.02049 -2.975
## prio                                                     0.02833  2.936
## relevel(educcat, "9th grade or less")10th to 11th grade  0.22716 -0.949
## relevel(educcat, "9th grade or less")12th grade          0.46543 -1.628
## relevel(educcat, "9th grade or less")Some college        1.00954 -1.225
##                                                         Pr(>|z|)   
## fin                                                      0.08077 . 
## age                                                      0.00293 **
## prio                                                     0.00332 **
## relevel(educcat, "9th grade or less")10th to 11th grade  0.34245   
## relevel(educcat, "9th grade or less")12th grade          0.10351   
## relevel(educcat, "9th grade or less")Some college        0.22072   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
##                                                         exp(coef) exp(-coef)
## fin                                                        0.7167     1.3952
## age                                                        0.9409     1.0628
## prio                                                       1.0868     0.9202
## relevel(educcat, "9th grade or less")10th to 11th grade    0.8060     1.2407
## relevel(educcat, "9th grade or less")12th grade            0.4687     2.1335
## relevel(educcat, "9th grade or less")Some college          0.2905     3.4428
##                                                         lower .95 upper .95
## fin                                                       0.49321    1.0416
## age                                                       0.90385    0.9794
## prio                                                      1.02805    1.1488
## relevel(educcat, "9th grade or less")10th to 11th grade   0.51641    1.2581
## relevel(educcat, "9th grade or less")12th grade           0.18825    1.1670
## relevel(educcat, "9th grade or less")Some college         0.04016    2.1009
## 
```

```
## Concordance= 0.655  (se = 0.026 )
## Likelihood ratio test= 33.46  on 6 df,   p=9e-06
## Wald test            = 30.37  on 6 df,   p=3e-05
## Score (logrank) test = 32  on 6 df,   p=2e-05
```

```r
summary(cox4)
```

```
## Call:
## coxph(formula = Surv(week, arrest) ~ fin + age + prio + educ,
##     data = df, ties = "breslow")
##
##   n= 432, number of events= 114
##
##          coef exp(coef) se(coef)      z Pr(>|z|)
## fin  -0.33086   0.71830  0.19029 -1.739  0.08208 .
## age  -0.06139   0.94045  0.02043 -3.005  0.00266 **
## prio  0.08177   1.08521  0.02832  2.887  0.00389 **
## educ -0.32458   0.72283  0.15246 -2.129  0.03326 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##      exp(coef) exp(-coef) lower .95 upper .95
## fin     0.7183     1.3922    0.4947    1.0430
## age     0.9405     1.0633    0.9035    0.9789
## prio    1.0852     0.9215    1.0266    1.1471
## educ    0.7228     1.3834    0.5361    0.9746
##
## Concordance= 0.655  (se = 0.026 )
## Likelihood ratio test= 33.05  on 4 df,   p=1e-06
## Wald test            = 30.73  on 4 df,   p=3e-06
## Score (logrank) test = 31.97  on 4 df,   p=2e-06
```

- Testing H0: The trend for the categorical EDUC variable is linear vs. H1: The trend for the categorical EDUC variable is not linear.
- Significance level: two sided $\alpha = 0.05$
- Test statistic: G $= -2$ [l(Reduced) $-$ l(Full)] = -2(-659.16-(-658.96)) = 0.4
- Decision rule: At $\alpha = 0.05$, reject H_0 if G $> \chi_2 = 5.99$
- Since G = 0.4<5.99, fail to reject H_0 with p = 0.8147
- We cannot reject the null hypothesis that effect of educcat on the log hazard is linear. That is, the categorical model does not provide a significantly better fit than the linear model(model with numeric educ), thus the linear representation is adequate (b/c it has less coefficient).

```r
anova(cox3,cox4)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(week, arrest)
##  Model 1: ~ fin + age + prio + relevel(educcat, "9th grade or less")
##  Model 2: ~ fin + age + prio + educ
##    loglik Chisq Df P(>|Chi|)
## 1 -658.96
## 2 -659.16  0.41  2    0.8147
```

If EDUC can be represented as a quantitative variable in the model, refit the model containing FIN, AGE, PRIO, and EDUC, where EDUC is not treated as a categorical variable.

```
cox5 = coxph(Surv(week, arrest) ~ fin + age + prio + educ,data=df, ties="breslow")
summary(cox5)
```

```
## Call:
## coxph(formula = Surv(week, arrest) ~ fin + age + prio + educ,
##     data = df, ties = "breslow")
##
##   n= 432, number of events= 114
##
##          coef exp(coef) se(coef)      z Pr(>|z|)
## fin  -0.33086   0.71830  0.19029 -1.739  0.08208 .
## age  -0.06139   0.94045  0.02043 -3.005  0.00266 **
## prio  0.08177   1.08521  0.02832  2.887  0.00389 **
## educ -0.32458   0.72283  0.15246 -2.129  0.03326 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##      exp(coef) exp(-coef) lower .95 upper .95
## fin     0.7183     1.3922    0.4947    1.0430
## age     0.9405     1.0633    0.9035    0.9789
## prio    1.0852     0.9215    1.0266    1.1471
## educ    0.7228     1.3834    0.5361    0.9746
##
## Concordance= 0.655  (se = 0.026 )
## Likelihood ratio test= 33.05  on 4 df,   p=1e-06
## Wald test            = 30.73  on 4 df,   p=3e-06
## Score (logrank) test = 31.97  on 4 df,   p=2e-06
```

d. Based on the model fit in part (c) (i.e., if you find that you can treat EDUC as a quantitative variable, then do so here), is FIN an important predictor of re-arrest at the $\alpha = 0.05$ -level in the model containing AGE, PRIO, and EDUC?

No, the fin variable has p-value $0.08208 > 0.05$, therefore it's not a significant redictor of rearrest at $\alpha = 0.05$ level.

Is there evidence that FIN is an important cofounder (looking for $\alpha >= 10\%$ change) that should be controlled for?

We remove the fin variable from the model, assess whether removal of the covariate has produced an "important" change (10-20%) in the coefficients of the variables remaining in the model.

```
cox6 = coxph(Surv(week, arrest) ~  age + prio + educ,data=df, ties="breslow")
summary(cox6)
```

```
## Call:
## coxph(formula = Surv(week, arrest) ~ age + prio + educ, data = df,
##     ties = "breslow")
##
##   n= 432, number of events= 114
##
##          coef exp(coef) se(coef)      z Pr(>|z|)
## age  -0.06348   0.93849  0.02035 -3.120  0.00181 **
## prio  0.07978   1.08305  0.02811  2.838  0.00454 **
## educ -0.33329   0.71656  0.15191 -2.194  0.02824 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##      exp(coef) exp(-coef) lower .95 upper .95
## age     0.9385     1.0655    0.9018    0.9767
## prio    1.0830     0.9233    1.0250    1.1444
## educ    0.7166     1.3955    0.5320    0.9651
##
## Concordance= 0.653  (se = 0.027 )
## Likelihood ratio test= 29.98  on 3 df,   p=1e-06
## Wald test            = 27.55  on 3 df,   p=5e-06
## Score (logrank) test = 28.7  on 3 df,   p=3e-06
```

To determine if it's a important confounder, we use (the coeff from the model with fin - coeff from the model w/o fin)/the coeff from the model with fin. That is,
For age: (-0.06139 - (-0.06348)) / (-0.06139) = -0.03404463,
For prio: (0.08177 - (0.07978)) / (0.08177) = 0.02433655,
For educ: (-0.32458 - (-0.33329)) / (-0.32458) = -0.02683468.
All the change is less than 10-20%. Therefore, it's not a significant confounder.

If not, remove FIN from the model. Interpret the hazard ratios from your final model.

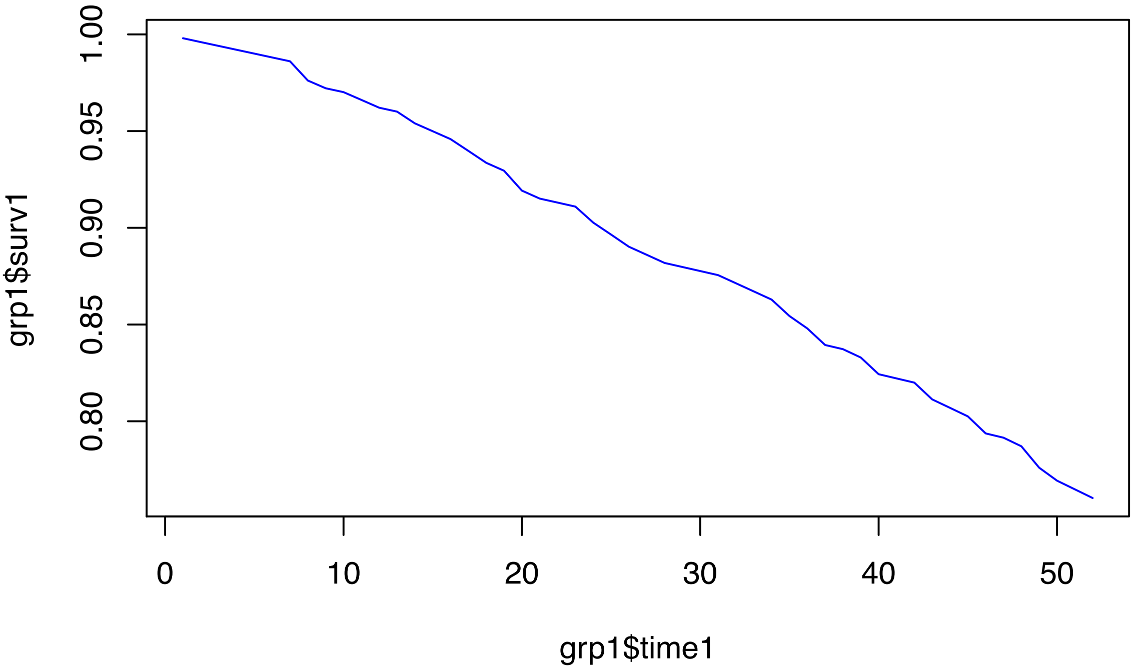
We conclude that fin is not important cofounder that should be controlled for. Please see *cox6* as the model without fin variable. Note that all of our variable are on a continuous scale. Therefore, we can interpret that The hazard of rearrest decreases by 6.15% for a 1-year increase in the inmate's age at time of release. The hazard of rearrest increases by 8.3% for a 1-count increase in the number of convictions prior to incarceration. The hazard of rearrest decreases by 28.34% for a 1-level increase in the inmate's highest level of completed schooling.

2. [20] In this question, you are asked to plot adjusted survival curves using the final model from question 1d.

a. Plot the overall adjusted survival curve using the mean values of the covariates in the model. Report the mean values used in constructing the plot.

mean value: age=24.62037,prio = 2.983796, educ=3.550926

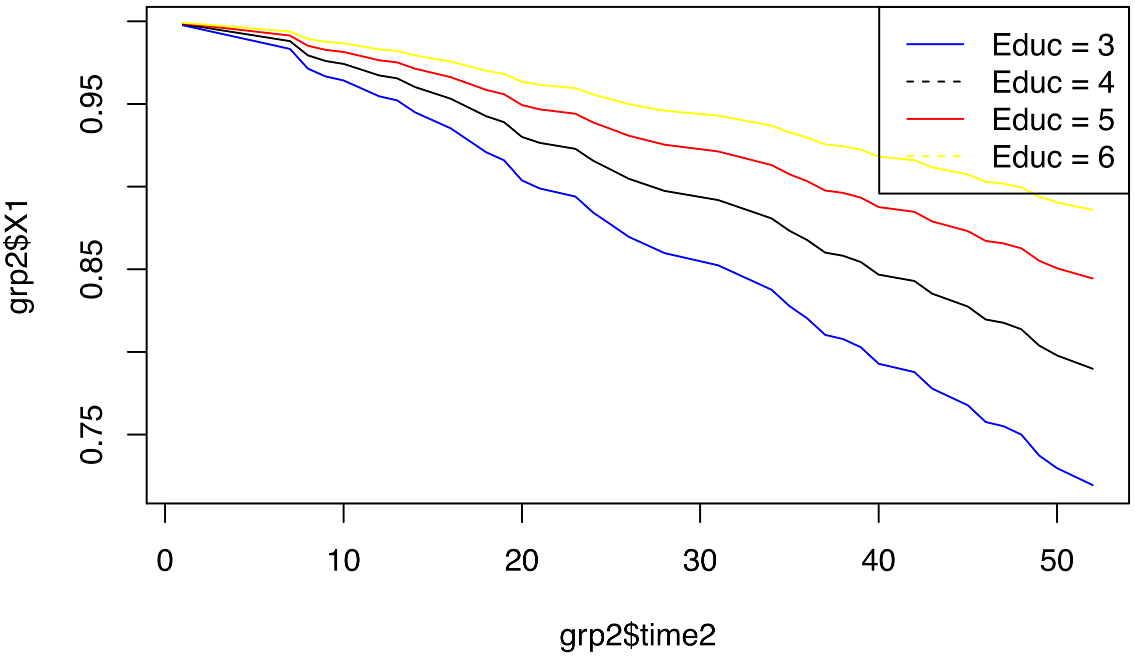
```
# Assessing Group variable controlling for the covariates
cox2.1 = coxph(Surv(week, arrest)~age + prio + educ, data=df, ties="breslow")
pred1 = survfit(cox2.1, newdata=data.frame(age=24.62037,prio = 2.983796, educ=3.550926)) #Estimated Cox

time1 = pred1$time
surv1 = pred1$surv

grp1 = data.frame(time1, surv1)
plot(grp1$time1, grp1$surv1, col="blue", type="l")
```



b. Plot the adjusted survival curves for EDUC = 3, 4, 5, and 6. Use the observed mean value of the other variables included in the model when estimating the adjusted survival curves for the four values of EDUC.

To compute the adjusted survival curve, we need to: - Stratify the data by treatment - Fit a PH model in each stratum, and then - Obtain adjusted survival probabilities using the overall mean logWBC in the estimated survival curve formula for each stratum

```
# Assessing Group variable controlling for the covariates
cox2.2 = coxph(Surv(week, arrest)~age + prio + educ, data=df, ties="breslow")
new = data.frame(age=c(24.62037,24.62037,24.62037,24.62037),prio = c(2.983796,2.983796,2.983796,2.98379
pred2 = survfit(cox2.2, data = df, newdata=new) #Estimated Cox survival curve

time2 = pred2$time
surv2 = pred2$surv
grp2 = data.frame(time2, surv2)
plot(grp2$time2, grp2$X1, col="blue", type="l")
lines(grp2$time2, grp2$X2, col="black", type="l")
lines(grp2$time2, grp2$X3, col="red", type="l")
lines(grp2$time2, grp2$X4, col="yellow", type="l")
legend('topright', legend=c("Educ = 3", "Educ = 4","Educ = 5","Educ = 6"), col=c("blue", "black","red",
```



Comment on the survival experiences observed for the different education levels in the context of this problem.

The four curves are roughly parallel, indicating that the PH assumption is satisfied for treatment after adjustment for Education. We can see that the inmate with the higher the education has higher probability of not being re-arrested.

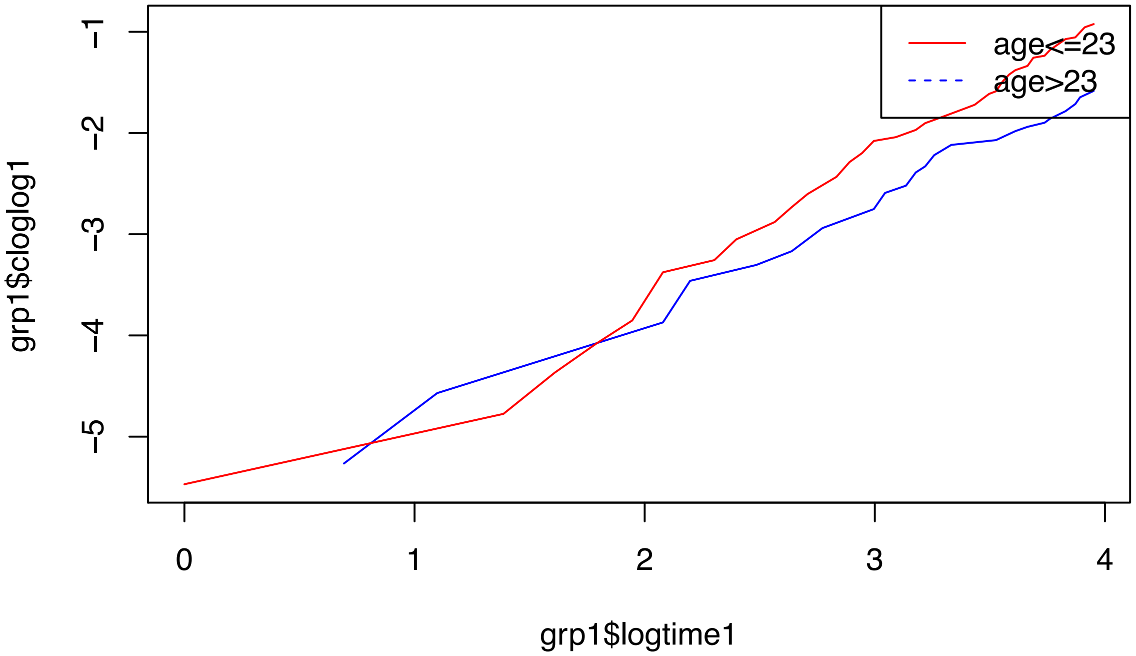
3. [40] In this question, you will check the proportional hazards (PH) assumption of AGE, PRIO, and EDUC in the final model from question 1d.

a. Use the log(-log(S(t)) plot using the Kaplan-Meier estimator for S(t) to assess the PH assumption for each variable. You will have to categorize any quantitative variables in order to use this graphical check. One option you could try is to dichotomize using each variable's median value as a cut-point (i.e., <= median, > median). Based on the figures, comment on the validity of the PH assumption for each variable.

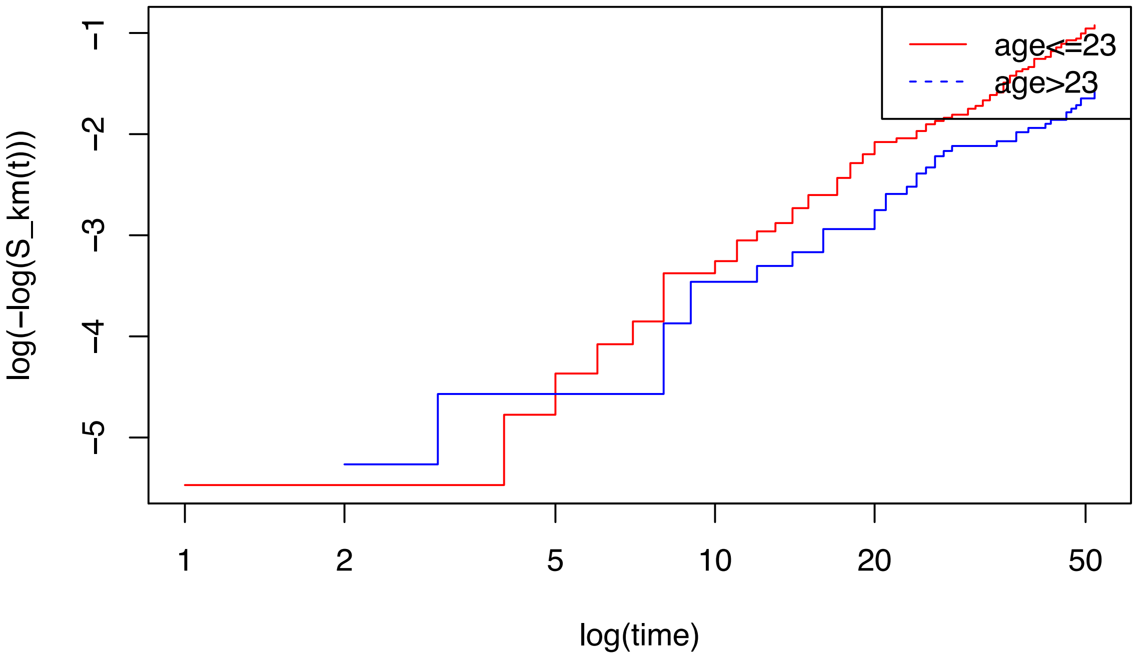
Note: Do not use the dichotomized versions of AGE, PRIO, and EDUC in the remainder of this question.

```
df$age_dicho = df[,ifelse(age>median(df$age),1,0)]
df$prio_dicho = df[,ifelse(prio>median(df$prio),1,0)]
df$educ_dicho = df[,ifelse(educ>median(df$educ),1,0)]

km1 = survfit(Surv(week, arrest) ~ 1, data = df[df$age_dicho==1,]) #KM by group
km0 = survfit(Surv(week, arrest) ~ 1, data = df[df$age_dicho==0,])

time1 = km1$time
logtime1 = log(time1) #log(time)
surv1 = km1$surv
cloglog1 = log(-log(surv1)) #log(-log(S_km(t)))
grp1 = data.frame(time1, logtime1, surv1, cloglog1)
grp1 = grp1[grp1$cloglog1!=Inf,] #In case survival curve ends at 0
```

```
time0 = km0$time
logtime0 = log(time0) #log(time)
surv0 = km0$surv
cloglog0 = log(-log(surv0)) #log(-log(S_km(t)))
grp0 = data.frame(time0, logtime0, surv0, cloglog0)
grp0 = grp0[grp0$cloglog0!=Inf,] #In case survival curve ends at 0
plot(grp1$logtime1, grp1$cloglog1, col="blue", type="l", xlim=c(min(grp0$logtime0,grp1$logtime1),max(grp
lines(grp0$logtime0, grp0$cloglog0, col="red", type="l")
legend('topright', legend=c("age<=23", "age>23"), col=c("red", "blue"), lty=1:2)
```

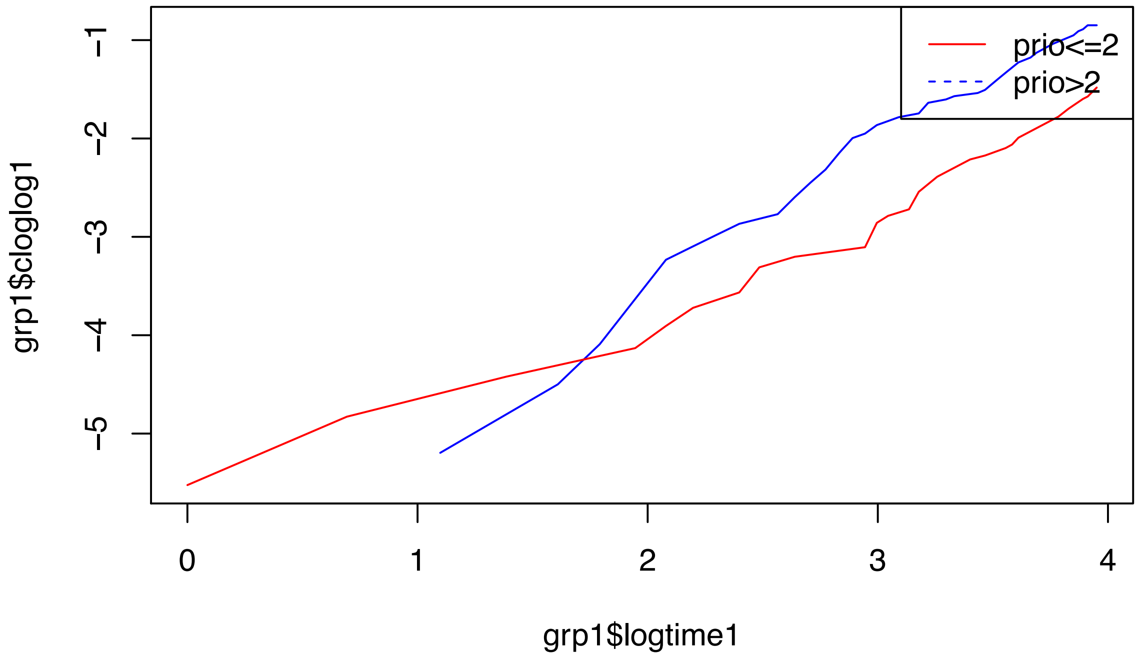


```
# Step function (built-in R)
plot(survfit(Surv(week, arrest) ~ age_dicho, data=df), col=c("red", "blue"), fun="cloglog", xlab="log(t
legend('topright', legend=c("age<=23", "age>23"), col=c("red", "blue"), lty=1:2)
```

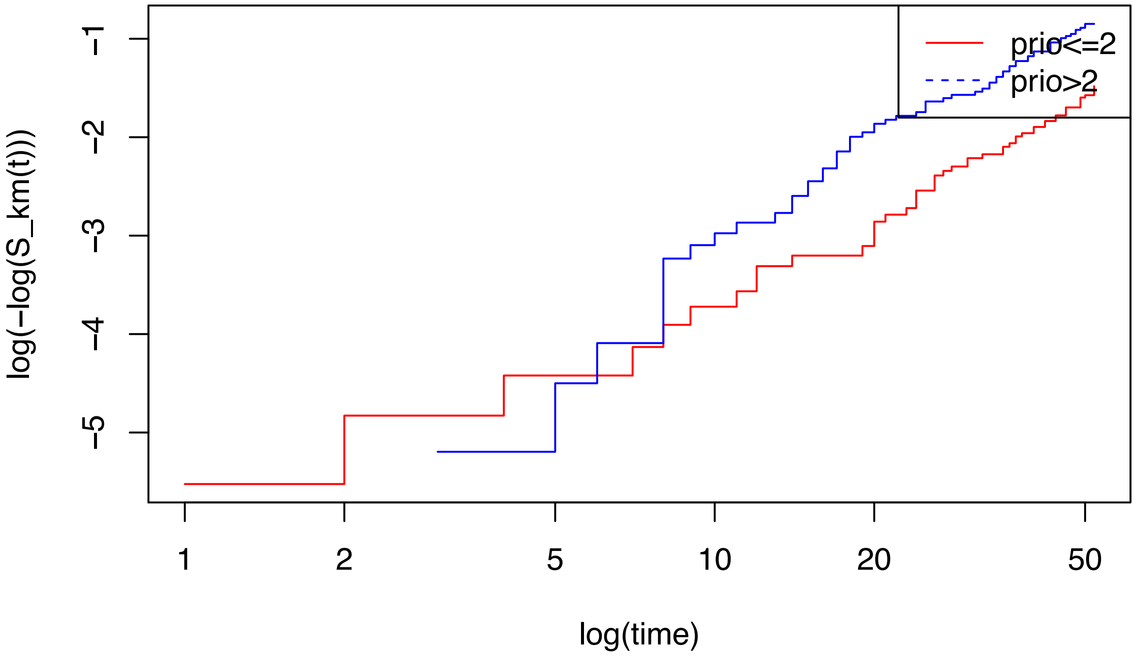


The two log-log survival plots clearly intersect, and are therefore nonparallel. Thus, age, when considered by itself, appears to violate the PH assumption.

```
km1.1 = survfit(Surv(week, arrest) ~ 1, data = df[df$prio_dicho==1,]) #KM by group
km1.0 = survfit(Surv(week, arrest) ~ 1, data = df[df$prio_dicho==0,])

time1 = km1.1$time
logtime1 = log(time1) #log(time)
surv1 = km1.1$surv
cloglog1 = log(-log(surv1)) #log(-log(S_km(t)))
grp1 = data.frame(time1, logtime1, surv1, cloglog1)
grp1 = grp1[grp1$cloglog1!=Inf,] #In case survival curve ends at 0

time0 = km1.0 $time
logtime0 = log(time0) #log(time)
surv0 = km1.0 $surv
cloglog0 = log(-log(surv0)) #log(-log(S_km(t)))
grp0 = data.frame(time0, logtime0, surv0, cloglog0)
grp0 = grp0[grp0$cloglog0!=Inf,] #In case survival curve ends at 0
plot(grp1$logtime1, grp1$cloglog1, col="blue", type="l", xlim=c(min(grp0$logtime0,grp1$logtime1),max(grp
lines(grp0$logtime0, grp0$cloglog0, col="red", type="l")
legend('topright', legend=c("prio<=2", "prio>2"), col=c("red", "blue"), lty=1:2)
```



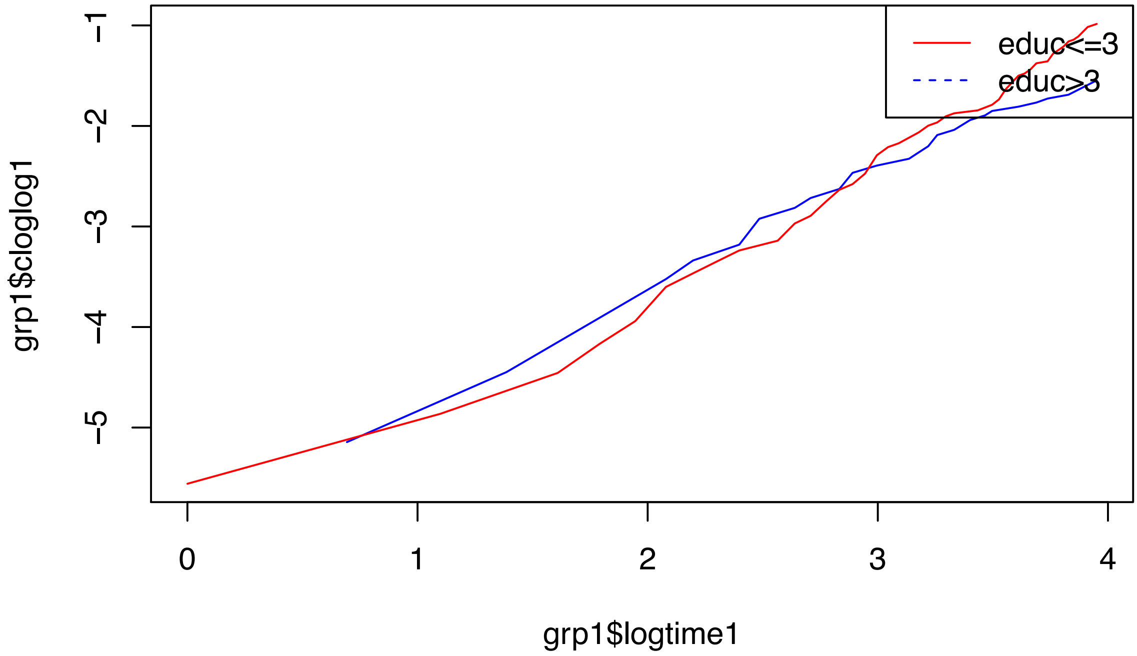
```
# Step function (built-in R)
plot(survfit(Surv(week, arrest) ~ prio_dicho, data=df), col=c("red", "blue"), fun="cloglog", xlab="log(
legend('topright', legend=c("prio<=2", "prio>2"), col=c("red", "blue"), lty=1:2)
```



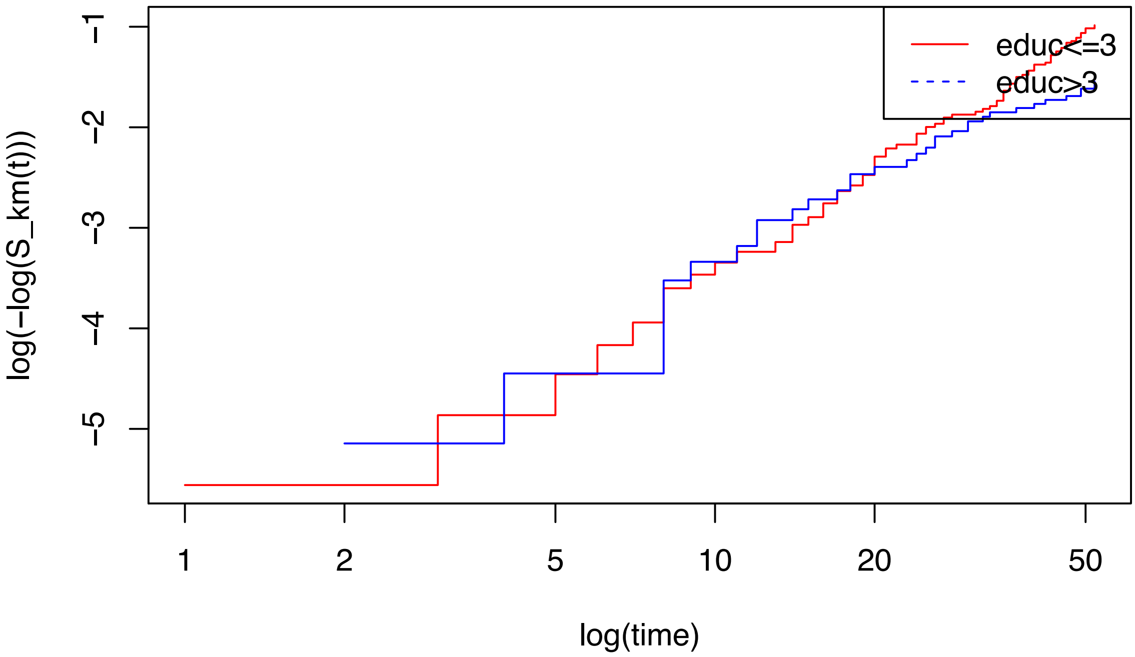
The two log-log survival plots clearly intersect, and are therefore nonparallel. Thus, prio, when considered by itself, appears to violate the PH assumption.

```
km1.1 = survfit(Surv(week, arrest) ~ 1, data = df[df$educ_dicho==1,]) #KM by group
km1.0 = survfit(Surv(week, arrest) ~ 1, data = df[df$educ_dicho==0,])

time1 = km1.1$time
logtime1 = log(time1) #log(time)
```

```
surv1 = km1.1$surv
cloglog1 = log(-log(surv1)) #log(-log(S_km(t)))
grp1 = data.frame(time1, logtime1, surv1, cloglog1)
grp1 = grp1[grp1$cloglog1!=Inf,] #In case survival curve ends at 0

time0 = km1.0 $time
logtime0 = log(time0) #log(time)
surv0 = km1.0 $surv
cloglog0 = log(-log(surv0)) #log(-log(S_km(t)))
grp0 = data.frame(time0, logtime0, surv0, cloglog0)
grp0 = grp0[grp0$cloglog0!=Inf,] #In case survival curve ends at 0
plot(grp1$logtime1, grp1$cloglog1, col="blue", type="l", xlim=c(min(grp0$logtime0,grp1$logtime1),max(gr
lines(grp0$logtime0, grp0$cloglog0, col="red", type="l")
legend('topright', legend=c("educ<=3", "educ>3"), col=c("red", "blue"), lty=1:2)
```



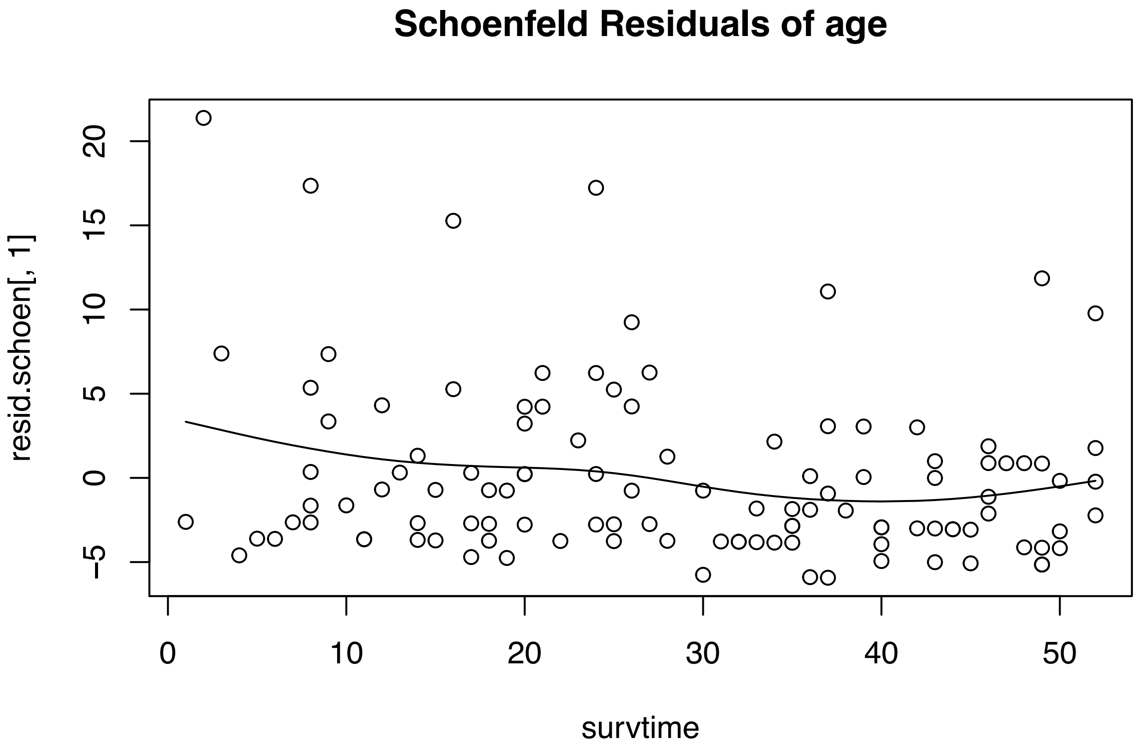
```
# Step function (built-in R)
plot(survfit(Surv(week, arrest) ~ educ_dicho, data=df), col=c("red", "blue"), fun="cloglog", xlab="log(
legend('topright', legend=c("educ<=3", "educ>3"), col=c("red", "blue"), lty=1:2)
```



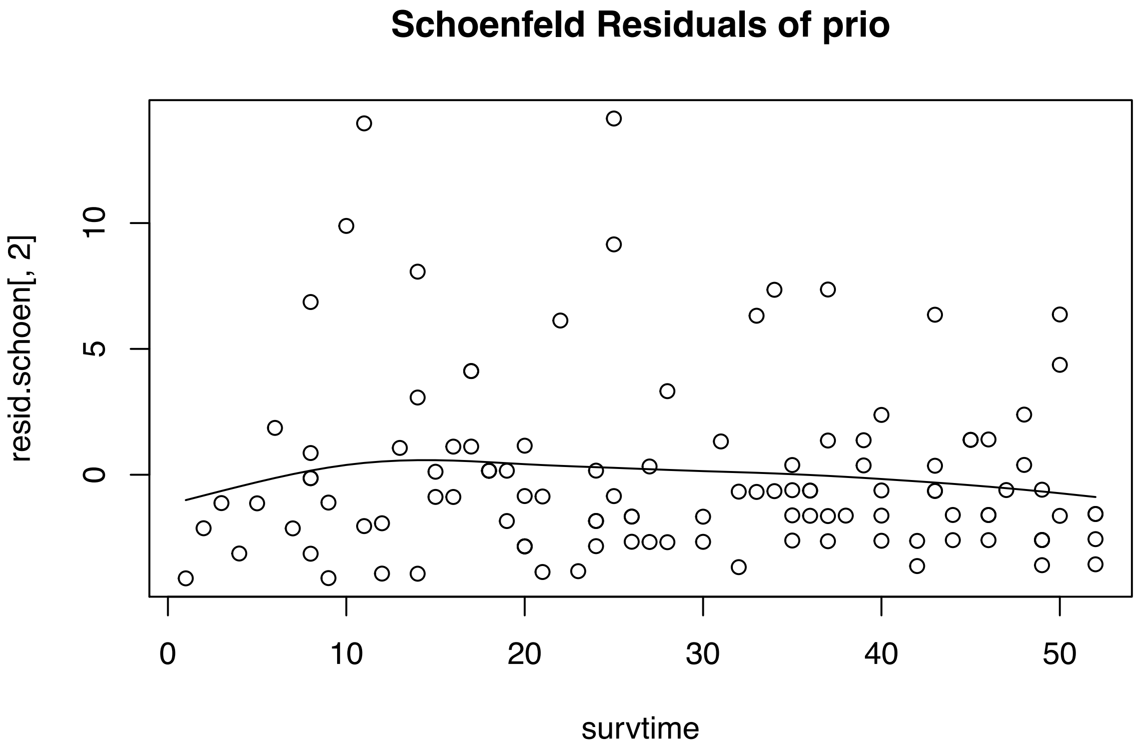
The two log-log survival plots clearly intersect, and are therefore nonparallel. Thus, educ, when considered by itself, appears to violate the PH assumption.

b. Use the Schoenfeld residuals for each covariate from the model to assess the PH assumption for each variable.

```
cox1 = coxph(Surv(week, arrest) ~  age + prio + educ,data=df, ties="breslow")
resid.schoen = residuals(cox1,type="schoenfeld")
survtime = as.numeric(rownames(resid.schoen))
plot(survtime, resid.schoen[,1], main="Schoenfeld Residuals of age")
smoothingSpline1 = smooth.spline(survtime, resid.schoen[,1], spar=0.85)
lines(smoothingSpline1)
```

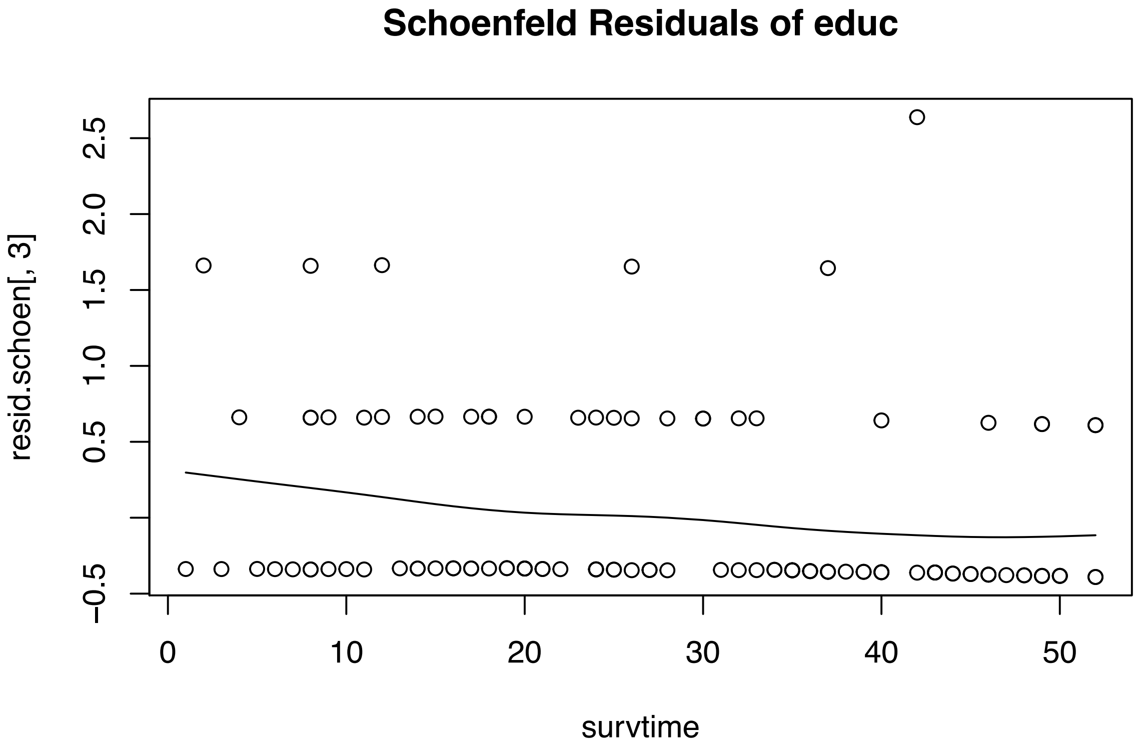


```
plot(survtime, resid.schoen[,2], main="Schoenfeld Residuals of prio")
smoothingSpline2 = smooth.spline(survtime, resid.schoen[,2], spar=0.85)
lines(smoothingSpline2)
```



```
plot(survtime, resid.schoen[,3], main="Schoenfeld Residuals of educ")
smoothingSpline2 = smooth.spline(survtime, resid.schoen[,3], spar=0.85)
```

```
lines(smoothingSpline2)
```



Comment on the plot of Schoenfeld residuals vs. the time variable (week)

Negative (non-zero) slope apparent when considering age and educ. Therefore, Proportional hazards assumption not supported for these variables. Flat slope for prio supports the proportional hazards assumption. (Note to myself: if there's pattern, that means the fit is not good bc our data iid. (interation/linear assumption). We mainly look at if the line (simulated) is flat or not.)

and perform a formal test of the correlation between the residuals and rank time (Note: Fully showing steps of hypothesis testing for the test of the correlation is not required here. Interpretation of the p-value of the test of correlation is sufficient.). Do the checks based on the Schoenfeld residuals suggest that any of the variables included in the model do not satisfy the PH assumption?

- Hypothesis : H_0: the correlation between the Schoenfeld residuals and ranked failure times = 0 H_1: the correlation between the Schoenfeld residuals and ranked failure times ≠ 0

```
# Cox model being evaluated
cox1 = coxph(Surv(week, arrest) ~ age + prio + educ, data=df, ties="breslow")
# Schoenfeld residuals of model
resid.schoen = residuals(cox1, type="schoenfeld")
survtime = as.numeric(rownames(resid.schoen))
# Correlation
survtime.rank = rank(survtime)
library(Hmisc)
```

```
## Loading required package: lattice
```

```
## Loading required package: Formula
## Loading required package: ggplot2

##
## Attaching package: 'Hmisc'

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

```
rcorr(as.matrix(data.frame(survtime.rank, resid.schoen[,1], resid.schoen[,2],resid.schoen[,3])))
```

```
##                   survtime.rank resid.schoen...1. resid.schoen...2.
## survtime.rank              1.00             -0.19             -0.09
## resid.schoen...1.         -0.19              1.00             -0.16
## resid.schoen...2.         -0.09             -0.16              1.00
## resid.schoen...3.         -0.18              0.29             -0.08
##                   resid.schoen...3.
## survtime.rank                 -0.18
## resid.schoen...1.              0.29
## resid.schoen...2.             -0.08
## resid.schoen...3.              1.00
##
## n= 114
##
##
## P
##                   survtime.rank resid.schoen...1. resid.schoen...2.
## survtime.rank                   0.0461            0.3684
## resid.schoen...1. 0.0461                          0.0992
## resid.schoen...2. 0.3684        0.0992
## resid.schoen...3. 0.0534        0.0019            0.4168
##                   resid.schoen...3.
## survtime.rank     0.0534
## resid.schoen...1. 0.0019
## resid.schoen...2. 0.4168
## resid.schoen...3.
```

Note that the p value for the test of the correlation coefficient = 0 are > 0.05 for prio and educ, we do not have evidence to reject the null hypothesis. Therefore, the PH assumption does not appear to be violated for prio and educ. For the age variable, the p value of the correlation between Schoenfeld residuals and ranked time is 0.0461 which is < 0.05. Therefore, we have evidence to reject the null hypothesis, suggesting that the PH assumption is violated for age.

c. Use an extended Cox model to test the PH assumption for AGE in a model containing PRIO and EDUC. Is there evidence to suggest that AGE violates the PH assumption in this model?

- Hypothesis H_0: the coefficient of the interaction of the age and some function of the week = 0 H_1: the coefficient of the interaction of the age and some function of the week $\neq 0$
- Significance level: two-sided $\alpha = 0.05$
- Test statistic: $W = coef/se(coef) = -1.936$, which is compared to +/-1.96.
- Decision rule: If$ W > z_.975 = 1.96$, we reject the null hypothesis.
- Statistical conclusion: Since |-1.936| < z_.975 = 1.96, we fail to reject with p value 0.05287 > 0.05.
- Conclusion: therefore, we don't have evidence to reject H0 and conclude that the interaction of age and time, age*time, is not significantly different from 0. Therefore, there's no evidence to suggest that age violates the PH assumption in the model after adjusting for educ and prio.

```r
# Convert data into counting process style, defined by vector of unique event times
cut.points = unique(df$week[df$arrest == 1])
df2 = survSplit(data = df, cut = cut.points, end = "week", start = "week0", event = "arrest")
df2 = df2[order(df2$ID),]
df2$agetime = df2$age*df2$week # Creating time-varying covariate to test
cox_extended1 = coxph(Surv(week0, week, arrest) ~ age + agetime + cluster(ID) + prio + educ, data=df2, 
summary(cox_extended1)
```

```
## Call:
## coxph(formula = Surv(week0, week, arrest) ~ age + agetime + prio +
##     educ, data = df2, method = "breslow", cluster = ID)
##
##   n= 18766, number of events= 114
##
##               coef exp(coef)  se(coef) robust se      z Pr(>|z|)
## age       0.024362  1.024661  0.039376  0.047340  0.515  0.60682
## agetime -0.003329  0.996676  0.001427  0.001720 -1.936  0.05287 .
## prio      0.080787  1.084140  0.028114  0.029896  2.702  0.00689 **
## educ     -0.331980  0.717502  0.151748  0.149611 -2.219  0.02649 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##         exp(coef) exp(-coef) lower .95 upper .95
## age        1.0247     0.9759    0.9339     1.124
## agetime    0.9967     1.0033    0.9933     1.000
## prio       1.0841     0.9224    1.0224     1.150
## educ       0.7175     1.3937    0.5351     0.962
##
## Concordance= 0.654  (se = 0.026 )
## Likelihood ratio test= 35.53  on 4 df,   p=4e-07
## Wald test            = 28.48  on 4 df,   p=1e-05
## Score (logrank) test = 31.99  on 4 df,   p=2e-06,   Robust = 31.41  p=3e-06
##
##   (Note: the likelihood ratio and score tests assume independence of
##      observations within a cluster, the Wald and robust score tests do not).
```

Note to myself: we can't use cox model w/o pH assumption. For testing PH, the stats test is most robust.

d. Report the equation of the final extended Cox model fit in part (c).

$h(t, X) = h_0(t) exp(0.024362 * age - 0.003329 * age * week + 0.080787 * prio - 0.331980 * educ)$

Interpret the hazard ratio associated with AGE from this model. Specifically, estimate the hazard ratio associated with a 1-year increase in age at 3 weeks, 20 weeks, and 50 weeks.

$\hat{HR}$ at 3 weeks = exp(0.024362-0.003329*3) = 1.014479;
$\hat{HR}$ at 20 weeks = exp(0.024362-0.003329*20) = 0.9586608;
$\hat{HR}$ at 50 weeks = exp(0.024362-0.003329*50) = 0.8675449.
The hazard of rearrest increases by 1.44% for a 1-year increase in the inmate's age at 3 weeks;
The hazard of rearrest decreases by 4.13% for a 1-year increase in the inmate's age at 20 weeks;
The hazard of rearrest decreases by 13.25% for a 1-year increase in the inmate's age at 50 weeks.

Comment on the differences that you observe.

The larger increase in inmate's age, the larger decrease the hazard of rearrest.