Homework 2

(P8110 Applied Regression II)

Mindy Tran (mnt2130)

Instructions

Submit your assignment as a single PDF document.

Label all graphics appropriately.

All calculations can be done in R/SAS/etc. except where specifically noted.

Include both your code and output inline (i.e. within each question and not all at the end of your document), but clearly state your response to the question. Code and output which is not accompanied by a description will not receive credit.

```r
library(tidyverse)
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.2 --
v ggplot2 3.4.1      v purrr   0.3.5
v tibble  3.1.8      v dplyr   1.0.10
v tidyr   1.2.1      v stringr 1.5.0
v readr   2.1.3      v forcats 0.5.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

```r
library(survival)
library(readxl)
library(knitr)
library(survminer)
```

```
Loading required package: ggpubr

Attaching package: 'survminer'

The following object is masked from 'package:survival':

    myeloma
```

```r
library(cmprsk)
```

Problem 1 (30 points)

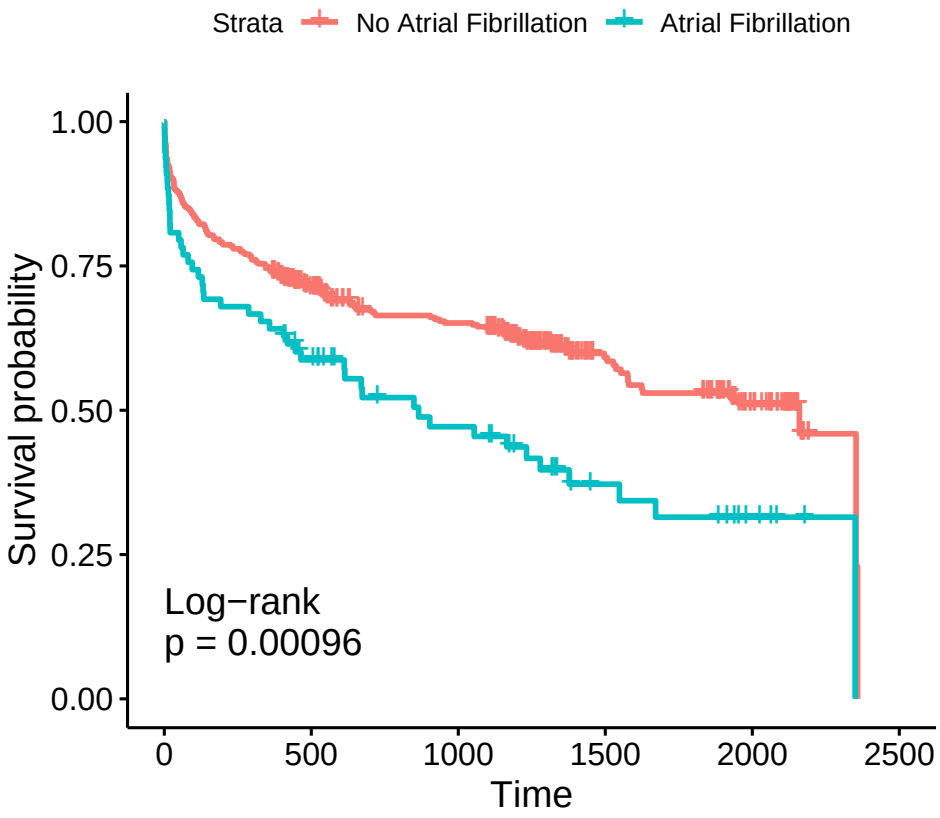
For this problem, utilize the `whas500` data set.

```r
# data import using readxl package
whas500 = read_xlsx("data/P8110_data.xlsx",sheet = "whas500")
```

1.1 (5 points)

Plot the Kaplan-Meier curve for survival (time to death), by atrial fibrillation (afib) group, using a single plot. Interpret the results and your expectation regarding whether a significant difference in survival exists between groups based on the plot.

```
# fits a Kaplan Meier curve for survival by atrial fibrilation group
fit1_1 = survfit(Surv(lenfol, fstat) ~ afb, data = whas500)

# creates a KM plot for both groups in a single plot
ggsurvplot(
fit1_1,
pval = TRUE,
pval.method = TRUE,
log.rank.weights = "1",
pval.coord = c(0, 0.1),
pval.method.coord = c(0, 0.17),
legend.labs = c("No Atrial Fibrillation", "Atrial Fibrillation"))
```



Survival appears to decreaese at a steady rate for both groups. After a couple days, the survival curves for each group (afib vs no afib) begins to show significant separation, suggesting that there is a difference between the two survival curves. To confirm, using the log-rank test and given that p= 0.00096, we have sufficient evidence to conclude that the two survival curves are different at the 0.05 level of significance.

1.2 (10 points)

Fit a univariate Cox proportional hazards (PH) model for survival, adjusting for afib. Estimate the corresponding hazard ratio and 95% confidence interval and interpret. Comment on the significance of afib with respect to survival.

```
#this code fits a Cox PH model for survival adjusting for afib and provides the hazard rat
cox_afb = coxph(Surv(lenfol, fstat) ~ afb, data = whas500)
cox_afb
```

```
Call:
coxph(formula = Surv(lenfol, fstat) ~ afb, data = whas500)

      coef exp(coef) se(coef)     z      p
afb 0.5397    1.7156   0.1654 3.263 0.0011

Likelihood ratio test=9.58  on 1 df, p=0.001962
n= 500, number of events= 215
```

```
summary(cox_afb)
```

```
Call:
coxph(formula = Surv(lenfol, fstat) ~ afb, data = whas500)

  n= 500, number of events= 215

      coef exp(coef) se(coef)     z Pr(>|z|)
afb 0.5397    1.7156   0.1654 3.263   0.0011 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    exp(coef) exp(-coef) lower .95 upper .95
afb     1.716     0.5829      1.24     2.373

Concordance= 0.537  (se = 0.014 )
Likelihood ratio test= 9.58  on 1 df,   p=0.002
Wald test            = 10.64  on 1 df,   p=0.001
Score (logrank) test = 10.9  on 1 df,   p=0.001
```

HR= 1.7156 (95% CI: 1.24, 2.373). The estimated death rate among patients with atrial fibrillation is 1.7156 times the estimated death rate among patients without atrial fibrillation. With 95% confidence, we estimate that the death rate among patients with atrial fibrillation could be as little as 1.24 times or as much as 2.373 times that of patients without atrial fibrillation. Patients without atrial fibrillation appear to have better survival outcomes compared to those with atrial fibrillation.

1.3 (10 points)

Fit a univariate Cox PH model for survival, adjusting for age. Estimate the corresponding hazard ratio and 95% confidence interval for a 5-year difference in age and interpret. Comment

on the significance of age with respect to survival.

```
#this code fits a Cox PH model for survival adjusting for age and provides the hazard rati
cox_age = coxph(Surv(lenfol, fstat) ~ age, data = whas500)
cox_age
```

```
Call:
coxph(formula = Surv(lenfol, fstat) ~ age, data = whas500)

        coef exp(coef) se(coef)     z      p
age 0.066339  1.068589 0.006079 10.91 <2e-16

Likelihood ratio test=142.1  on 1 df, p=< 2.2e-16
n= 500, number of events= 215
```

```
summary(cox_age)
```

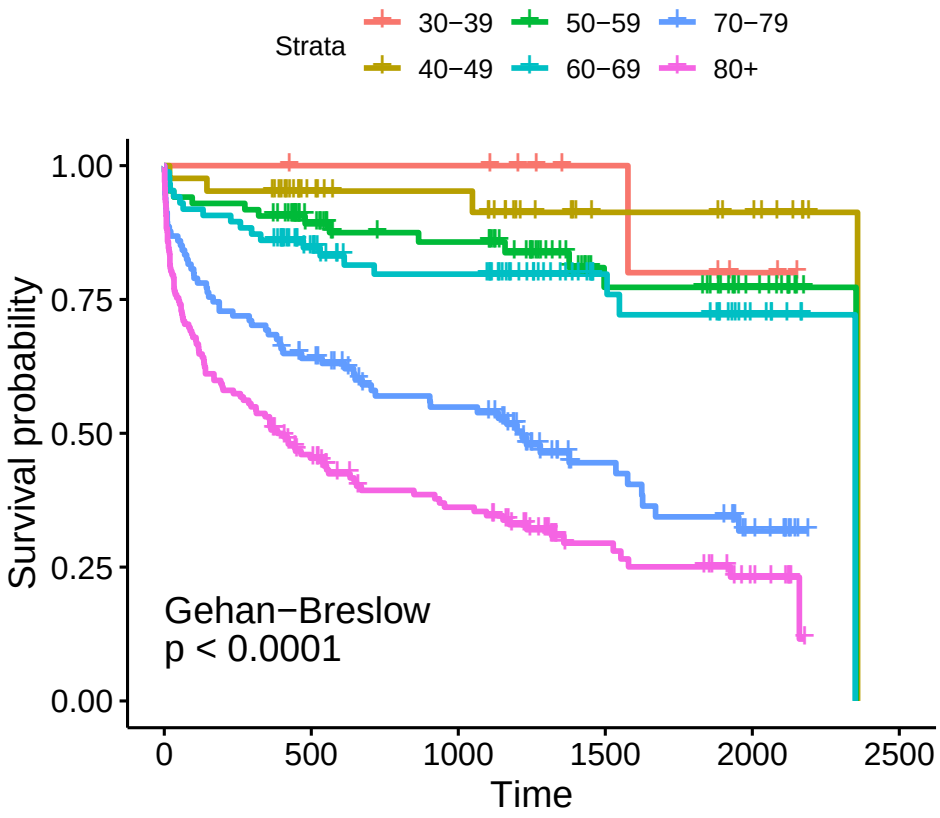
```
Call:
coxph(formula = Surv(lenfol, fstat) ~ age, data = whas500)

  n= 500, number of events= 215

        coef exp(coef) se(coef)     z Pr(>|z|)
age 0.066339  1.068589 0.006079 10.91   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    exp(coef) exp(-coef) lower .95 upper .95
age     1.069     0.9358     1.056     1.081

Concordance= 0.731  (se = 0.018 )
Likelihood ratio test= 142.1  on 1 df,   p=<2e-16
Wald test            = 119.1  on 1 df,   p=<2e-16
Score (logrank) test = 126.6  on 1 df,   p=<2e-16
```

HR= 1.068589 (95% CI: 1.056, 1.081). The estimated death rate increases by about 6.86% for every one-year increase in age. With 95% confidence, we estimate that a one year increase in age is associated with as little as 5.6% increase or as much as 8.1% increase in the risk of death.

1.4 (5 points)

Come up with a visualization for the difference in survival due to age. Describe and interpret your plot.

```
# this code breaks up the continuous age variable into 6 different age groups
whas500$age_group = cut(whas500$age, breaks = c(30,39,49,59,69,79,Inf), include.lowest = T
labels = c("30-39", "40-49", "50-59", "60-69","70-79", "80+"))

# this code fits a survival estimate for survival by age group and creates a KM plot for e
fit_age_group = survfit(Surv(lenfol, fstat) ~ age_group, data = whas500)

ggsurvplot(
fit_age_group,
pval = TRUE,
pval.method = TRUE,
log.rank.weights = "n",
pval.coord = c(0, 0.1),
pval.method.coord = c(0, 0.17),
legend.labs = c("30-39", "40-49", "50-59", "60-69","70-79", "80+"))
```



The plot above breaks up patients by age group and plots a KM survival curve for each age group.

From this plot, we observe that patients who are 80 years old or older experience the most rapid decline in survival probability, followed by those in the 70-79 age group. Those in the 30-39 age group appear to have the best survival probability, which drops at around 1500 days. It also appears that no additional deaths occur in that group after that time. Those who are in the younger age groups tend to have better survival probability over time compared to those in older age groups. There is a distinct gap in survival functions for those in the 30-39, 40-49, 50-59, 60-69 age group (as these survival functions are more clustered together) compared to those who are in the 70-79 age group and those who are in the 80+ age group. Since the survival curves appear to cross, using the Gehan-Breslow test, we observe that at least one age group has a different survival function than the others groups at the 0.05 level of significance ($p<0.0001$).

Problem 2 (25 points)

For this problem, utilize the `whas500` data set.

2.1 (10 points)

Assume the primary covariate of interest is afib history. Assess whether or not age is a confounder with respect to afib. Interpret your findings.

```
#this code generates the base model with afb as the primary covariate of interest
cox_afb = coxph(Surv(lenfol, fstat) ~ afb, data = whas500)
cox_afb
```

```
Call:
coxph(formula = Surv(lenfol, fstat) ~ afb, data = whas500)

      coef exp(coef) se(coef)     z      p
afb 0.5397    1.7156   0.1654 3.263 0.0011

Likelihood ratio test=9.58  on 1 df, p=0.001962
n= 500, number of events= 215
```

```
#this code generates the adjusted model with age as another variable
cox_adjusted = coxph(Surv(lenfol, fstat) ~ afb + age, data = whas500)
cox_adjusted
```

```
Call:
coxph(formula = Surv(lenfol, fstat) ~ afb + age, data = whas500)

        coef exp(coef) se(coef)      z      p
afb 0.337009  1.400751 0.165941  2.031 0.0423
age 0.066150  1.068387 0.006166 10.728 <2e-16

Likelihood ratio test=146  on 2 df, p=< 2.2e-16
n= 500, number of events= 215
```

```
# This code calculates the percent change in afib coefficient
100 * (coef(cox_afb) - coef(cox_adjusted)["afb"] ) / coef(cox_adjusted)["afb"]
```

```
     afb
60.15619
```

From the output above, we observe that the coefficient for afib from the crude model (Beta=0.5397) is different from the coefficient for afib from the adjusted model (Beta=0.337009). From the percent change in afib coefficient, which is calculated to be ~60.156%, this value is greater than the threshold set at an absolute value of 20% since no clinical guidelines were given. Since the value is greater than the threshold, we have evidence to suggest that age is a confounder for the relationship between afb and survival.

2.2 (10 points)

Assume the primary covariate of interest is afib history. Assess whether or not age is an effect modifier with respect to afib. Interpret your findings.

```
#test for effect modification of age
cox_em = coxph(Surv(lenfol, fstat) ~ afb * age, data = whas500)
cox_em
```

```
Call:
coxph(formula = Surv(lenfol, fstat) ~ afb * age, data = whas500)

             coef exp(coef)  se(coef)      z      p
afb     -0.964098  0.381327  1.577273 -0.611  0.541
age      0.064336  1.066451  0.006503  9.894 <2e-16
afb:age  0.016491  1.016628  0.019778  0.834  0.404

Likelihood ratio test=146.7  on 3 df, p=< 2.2e-16
n= 500, number of events= 215
```

```
summary(cox_em)
```

```
Call:
coxph(formula = Surv(lenfol, fstat) ~ afb * age, data = whas500)

  n= 500, number of events= 215

             coef exp(coef)  se(coef)      z Pr(>|z|)
```

```
afb     -0.964098  0.381327  1.577273 -0.611    0.541
age      0.064336  1.066451  0.006503  9.894   <2e-16 ***
afb:age  0.016491  1.016628  0.019778  0.834    0.404
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

        exp(coef) exp(-coef) lower .95 upper .95
afb        0.3813     2.6224   0.01733     8.392
age        1.0665     0.9377   1.05295     1.080
afb:age    1.0166     0.9836   0.97797     1.057

Concordance= 0.729  (se = 0.018 )
Likelihood ratio test= 146.7  on 3 df,   p=<2e-16
Wald test            = 122.1  on 3 df,   p=<2e-16
Score (logrank) test = 135.7  on 3 df,   p=<2e-16
```

From the output above, we see that the interaction term between afb and age yields a coefficient of 0.016491. The p-value of 0.404 is greater than 0.05, indicating that we do not have sufficient evidence to reject the null hypothesis that the coefficient of the interaction term is 0, thus our interaction coefficient is not significantly different from 0 at the 0.05 level of significance. Since the interaction term coefficient is not significantly different from 0, that afib does not vary with age (Age is not an effect modifier with respect to afib).

2.3 (5 points)

Based on the previous results, fit the most appropriate model with respect to including age, afib, or both. Interpret the model, including the resulting hazard ratio(s).

```r
#this code fits the Cox PH model with afb controlling for age
cox_adjusted = coxph(Surv(lenfol, fstat) ~ afb + age, data = whas500)

#this code generates the model outputs
cox_adjusted
```

```
Call:
coxph(formula = Surv(lenfol, fstat) ~ afb + age, data = whas500)

        coef exp(coef) se(coef)      z      p
afb 0.337009  1.400751 0.165941  2.031 0.0423
age 0.066150  1.068387 0.006166 10.728 <2e-16
```

```
Likelihood ratio test=146  on 2 df, p=< 2.2e-16
n= 500, number of events= 215
```

```
summary(cox_adjusted)
```

```
Call:
coxph(formula = Surv(lenfol, fstat) ~ afb + age, data = whas500)

  n= 500, number of events= 215

        coef exp(coef) se(coef)      z Pr(>|z|)
afb 0.337009  1.400751 0.165941  2.031   0.0423 *
age 0.066150  1.068387 0.006166 10.728   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    exp(coef) exp(-coef) lower .95 upper .95
afb     1.401     0.7139     1.012     1.939
age     1.068     0.9360     1.056     1.081

Concordance= 0.729  (se = 0.018 )
Likelihood ratio test= 146  on 2 df,   p=<2e-16
Wald test            = 120  on 2 df,   p=<2e-16
Score (logrank) test = 129.3  on 2 df,   p=<2e-16
```

Since age is shown to be a confounder, we want to include it in our model along with afib to control for age when estimating the association between afib and survival.

From the output generated above from our selected model, HR= 1.400751 (1.012, 1.939) when controlling for age. Thus, the estimated death rate among patients with atrial fibrillation is ~1.40 times the estimated death rate among patients without atrial fibrillation, when controlling for age. With 95% confidence, we estimate that the death rate among patients with atrial fibrillation could be as little as 1.012 times or as much as 1.939 times that of patients without atrial fibrillation when controlling for age.

Problem 3 (10 points)

For this problem, utilize the `whas500` data set.

Use the partial likelihood ratio test to compare a model containing age, BMI, and afib history to a model containing only afib history. Provide the test statistics, p-value, and interpret the results.

What assumption must be true regarding the two models in order to apply the partial likelihood ratio test?

```
# this code creates the first model containing age, BMI, and afib history
cox_model_1 = coxph(Surv(lenfol, fstat) ~ afb + age + bmi, data = whas500)

cox_model_1
```

```
Call:
coxph(formula = Surv(lenfol, fstat) ~ afb + age + bmi, data = whas500)

         coef exp(coef)  se(coef)      z      p
afb  0.302742  1.353565  0.166365  1.820 0.0688
age  0.060147  1.061993  0.006516  9.231 <2e-16
bmi -0.039311  0.961452  0.015360 -2.559 0.0105

Likelihood ratio test=152.7  on 3 df, p=< 2.2e-16
n= 500, number of events= 215
```

```
# a model containing afib was already generated for a previous problem, so we will use tha
cox_model_2 = cox_afb

cox_model_2
```

```
Call:
coxph(formula = Surv(lenfol, fstat) ~ afb, data = whas500)

      coef exp(coef) se(coef)     z      p
afb 0.5397    1.7156   0.1654 3.263 0.0011

Likelihood ratio test=9.58  on 1 df, p=0.001962
n= 500, number of events= 215
```

```r
# this compares the two models
anova(cox_model_1,cox_model_2)
```

```
Analysis of Deviance Table
 Cox model: response is  Surv(lenfol, fstat)
 Model 1: ~ afb + age + bmi
 Model 2: ~ afb
   loglik  Chisq Df Pr(>|Chi|)
1 -1151.0
2 -1222.5 143.13  2  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the generated output above, the test statistic= Chi-Square= 143.13 where p= 2.2 e^-16 at 2 degrees of freedom. Since we have a very small p-value that is less than 0.05, we have sufficient evidence to reject the null hypothesis and conclude that the more complex model (containing age, BMI and afib) fits the data significantly better than the simpler model (only containing afib) at the 0.05 level of significance. Thus, age, BMI, and afib history all have a significant effect on survival outcomes

To apply the partial likelihood the two models must meet the assumption that the two models are nested such that the simpler model is a subset of the more complex model and both models must use the same dataset and the same set of variables where the proportional hazard assumptions must be met for both models and they are fitted using the same method.

Problem 4 (35 points)

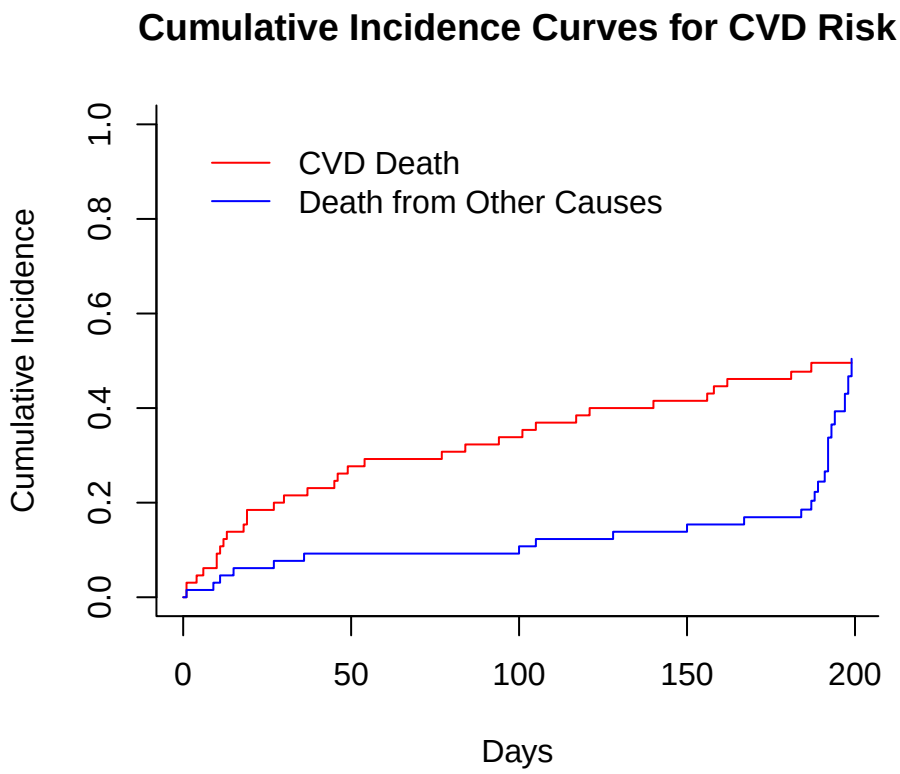
For this problem, utilize the `cvdrisk` data set. Treat CVD death as the primary event of interest, and death due to other causes as a competing risk.

```
# data import using readxl package
cvdrisk = read_xlsx("data/P8110_data.xlsx",sheet = "cvdrisk")
```

4.1 (5 points)

Plot the cumulative incidence of CVD death and death due to other causes.

```
# this code fits a cumulative incidence function to the CVD dataset
cvd_fit <- cuminc(
ftime = cvdrisk$time,
fstatus = cvdrisk$ev_typ)

# this code generates a plot of cumulative incidence of CVD death and death due to other c

plot(
cvd_fit,
xlab = "Days",
ylab = "Cumulative Incidence",
main = "Cumulative Incidence Curves for CVD Risk",
curvlab = c("CVD Death", "Death from Other Causes"),
color = c("red", "blue"),
lty = 1
)
```



This is a plot of the cumulative incidence curves for CVD deaths and deaths due to other causes using the CVDRisk dataset.

4.2 (10 points)

Estimate the cumulative incidence of CVD death and death due to other causes at 90 days, as well as the corresponding 95% confidence intervals.

```
# this uses the cumulative risk function to get an estimate of the cumulative incidence of
fit_est = timepoints(cvd_fit, times = c(90))
fit_est
```

```
$est
            90
1 1 0.32307692
```

```
1 2 0.09230769

$var
             90
1 1 0.003435636
1 2 0.001313384
```

```r
# this generates the 95% CI for the estimate at 90 days
est_ci =
tibble(
outcome = c("CVD", "other"),
est = fit_est$est[, 1],
var = fit_est$var[, 1],
ci_lower = est - qnorm(0.975) * sqrt(var),
ci_upper = est + qnorm(0.975) * sqrt(var)
)

est_ci
```

```
# A tibble: 2 x 5
  outcome    est     var ci_lower ci_upper
  <chr>    <dbl>   <dbl>    <dbl>    <dbl>
1 CVD     0.323  0.00344   0.208     0.438
2 other   0.0923 0.00131   0.0213    0.163
```

From the output above, the cumulative incidence of CVD death at 90 days is equal to 0.323 (95% CI: 0.208, 0.438) . The cumulative incidence of death due to other causes at 90 days is equal to 0.092 (95% CI: 0.0213, 0.163).

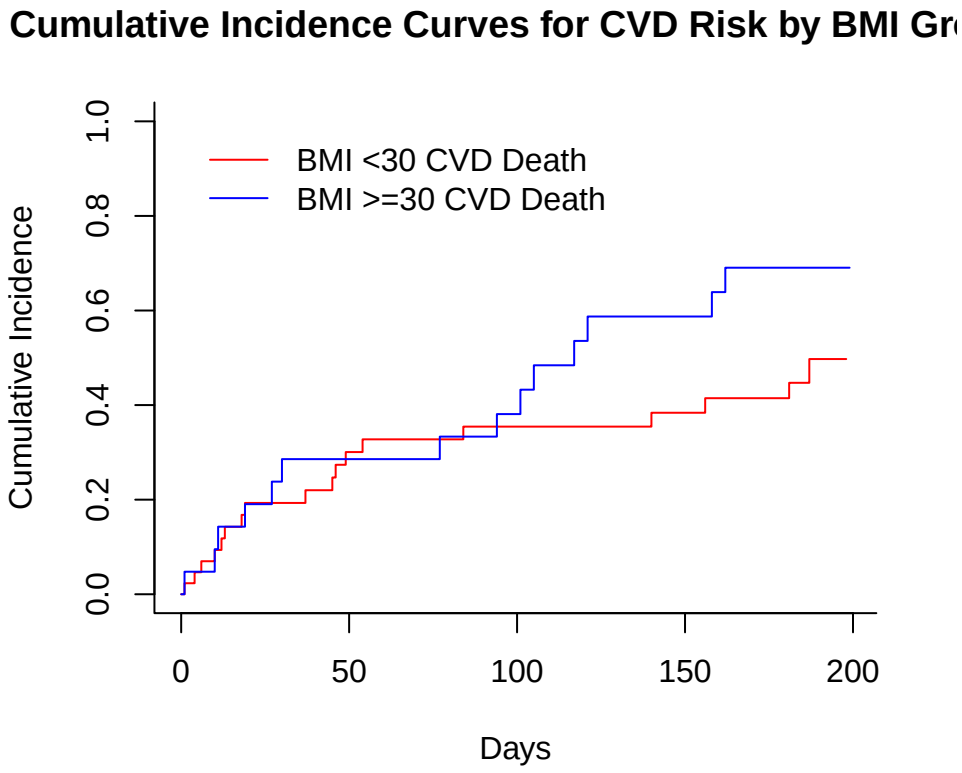
4.3 (10 points)

Plot the cumulative incidence of CVD death among individuals with a BMI < 30 vs. BMI ≥ 30.

Test for a difference in the cumulative incidence of CVD death between the BMI groups and interpret your results.

```r
#this code creates a new variable bmi group which is a factor variable where 0 = BMI <30 a
cvdrisk$bmi_group =ifelse(cvdrisk$bmi < 30, 0, 1)
```

```r
cvdrisk$bmi_group = factor(cvdrisk$bmi_group, levels = c(0, 1))

# this code fits a cumulative incidence function for only the outcome CVD deaths for each

bmi_fit <- cuminc(
ftime = cvdrisk$time,
fstatus = cvdrisk$ev_typ == 1,
group = cvdrisk$bmi_group
)



# this plots the cumulative incidence of CVD death among individuals with a BMI< 30vs. BMI

plot(
bmi_fit,
xlab = "Days",
ylab = "Cumulative Incidence",
main = "Cumulative Incidence Curves for CVD Risk by BMI Group",
curvlab = c("BMI <30 CVD Death", "BMI >=30 CVD Death"),
color = c("red", "blue"),
lty = c(1, 1)
)
```




```
# this code generates the output for cumulative incidence function
bmi_fit
```

```
Tests:
         stat        pv df
TRUE 1.672686 0.1958992  1
Estimates and Variances:
$est
               50       100       150
0 TRUE 0.3006413 0.3544381 0.3837818
1 TRUE 0.2857143 0.3809524 0.5873016

$var
                 50          100          150
0 TRUE 0.005515937 0.006090295 0.006410269
1 TRUE 0.010289492 0.011951330 0.013186470
```

For CVD Deaths: With a test statistic of 1.67 and a p-value of 0.196, we have insufficient evidence to reject the null hypothesis and conclude that there is no significant difference between cumulative incidence of CVD deaths between those with BMI <30 and those with BMI greater than or equal to 30 at the 0.05 level of significance.

4.4 (10 points)

Fit a cause-specific hazard Cox model for the risk of CVD death, adjusting for BMI group. Estimate the hazard ratio and corresponding 95% confidence interval. Interpret the results and compare with those of **4.3**.

```
#Cause specific hazard cox model for death from CVD including bmi_group
cause_mdl = coxph(
Surv(time, ev_typ == 1) ~ bmi_group,
data = cvdrisk
)

summary(cause_mdl)
```

```
Call:
coxph(formula = Surv(time, ev_typ == 1) ~ bmi_group, data = cvdrisk)

  n= 65, number of events= 32

             coef exp(coef) se(coef)     z Pr(>|z|)
bmi_group1 0.4628    1.5885   0.3572 1.296    0.195

           exp(coef) exp(-coef) lower .95 upper .95
bmi_group1     1.589     0.6295    0.7887     3.199

Concordance= 0.545  (se = 0.044 )
Likelihood ratio test= 1.63  on 1 df,   p=0.2
Wald test            = 1.68  on 1 df,   p=0.2
Score (logrank) test = 1.71  on 1 df,   p=0.2
```

```
#Cause specific hazard cox model for death from other causes including bmi_group
bmi_other_mdl = coxph(
Surv(time, ev_typ == 2) ~ bmi_group,
data = cvdrisk
```

```
)
summary(bmi_other_mdl)
```

```
Call:
coxph(formula = Surv(time, ev_typ == 2) ~ bmi_group, data = cvdrisk)

  n= 65, number of events= 24

              coef exp(coef) se(coef)      z Pr(>|z|)
bmi_group1 -0.8014    0.4487   0.5531 -1.449    0.147

           exp(coef) exp(-coef) lower .95 upper .95
bmi_group1    0.4487      2.229    0.1518     1.327

Concordance= 0.607  (se = 0.041 )
Likelihood ratio test= 2.46  on 1 df,   p=0.1
Wald test            = 2.1  on 1 df,   p=0.1
Score (logrank) test = 2.21  on 1 df,   p=0.1
```

For CVD Deaths: HR= 1.5885. The estimated death rate caused by CVD among patients with BMI >= 30 is 1.5885 (95% CI: 0.7887,3.199) times the estimated death rate caused by CVD among patients with BMI <30. With 95% confidence, we estimate that the death rate caused by CVD among patients with BMI >= 30 could be as little as 0.7887 times or as much as 3.199 times that of patients with BMI >= 30

For Other Cause Deaths: HR= 0.4487. The estimated death rate by other causes among patients with BMI >= 30 is 0.4487 (95% CI: 0.1518, 1.327) times the estimated death rate by other causes among patients with BMI <30. With 95% confidence, we estimate that the death rate by other causes among patients with BMI >= 30 could be as little as 0.1518 times or as much as 1.3277 times that of patients with BMI >= 30

Note: For both outcomes, the 95% CI includes the null value of 1, thus the HRs could be not statistically significant and the p-values of the coefficients are > 0.05, which could indicate that they are not statistically significantly different from 0.

Compared to the output from 4.3, we got similar results. For the cause specific model, if we look at the log rank test, we get a test statistic of 1.71 with p= 0.2, since p > 0.05, we have insufficient evidence to reject the null and conclude that there is no significant difference between the cumulative incidence of CVD deaths between those with BMI <30 and those with BMI greater than or equal to 30 at the 0.05 level of significance. This is aligned with what was observed in 4.3