qbs121_final_exam_gibran

Gibran Erlangga

3/9/2022

Data Analysis 1

The data set "skindata" is on the Canvas site. The outcome variable Y is a count of the number of new skin cancers per year. The categorical variable Treatment is coded 1=beta-carotene, 0=placebo. The variable Year denotes the year of follow-up. The categorical variable Gender is coded 1=male, 0=female. The categorical variable Skin denotes the skin susceptibility and is coded 1=burns easily, 0=otherwise. The variable Exposure is a count of the number of previous skin cancers. The variable Age is the age (in years) of each subject at randomization.

```
skin_data <- read.csv('skindata.csv')
head(skin_data, 3)
```

```
##       ID Center Age Skin Gender Exposure Y Treatment Year
## 1 100034      1  51    1      1        4 0         0    1
## 2 100034      1  51    1      1        4 1         0    2
## 3 100034      1  51    1      1        4 1         0    3
```

Variable List: ID, Center, Age, Skin, Gender, Exposure, Y, Treatment, Year.

Reference: Greenberg, E.R., Baron, et. Al. (1990). A clinical trial of beta carotene to prevent basal-cell and squamous-cell cancers of the skin. New England Journal of Medicine, 323, 789-795.

1. For these data, an "intention to treat" (ITT) analysis which only looks at Treatment as a factor, while adjusting for clustering and longitudinal structure is conventional

a. Fit a generalized linear mixed model (glmer), Poisson family, with Y as an outcome, a log link function, an effect for Treatment, Year as a continuous variable, and a random intercept for the individual. Write equations to specify this model and state the assumptions.

```
library(lme4)
```

```
## Loading required package: Matrix
```

```
## Warning: package 'Matrix' was built under R version 4.1.1
```

```
model <- glmer(Y ~ Treatment + Year + (1 | ID), family=poisson(link="log"), data=skin_data)
summary(model)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula: Y ~ Treatment + Year + (1 | ID)
##    Data: skin_data
##
##      AIC      BIC   logLik deviance df.resid
##   8426.5   8453.9  -4209.2   8418.5     7077
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.7067 -0.3827 -0.2390 -0.2264  6.5960
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  ID     (Intercept) 2.189    1.48
## Number of obs: 7081, groups:  ID, 1683
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.46461    0.09886 -24.930   <2e-16 ***
## Treatment    0.17309    0.09978   1.735   0.0828 .
## Year         0.01827    0.01792   1.020   0.3080
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr) Trtmnt
## Treatment -0.538
## Year      -0.504  0.010
```

Assumptions: I assume that there is a different trend happened for every individual, so we set the individual as the random intercept of the model.

b. Give the estimated variance of the random effects and show a histogram of the estimated random effects empirical Bayes estimates. With glmer, EB estimation requires the package merTools. The function REextract() extracts the EB estimates. Does the model seem reasonable?

```
library(merTools)
```

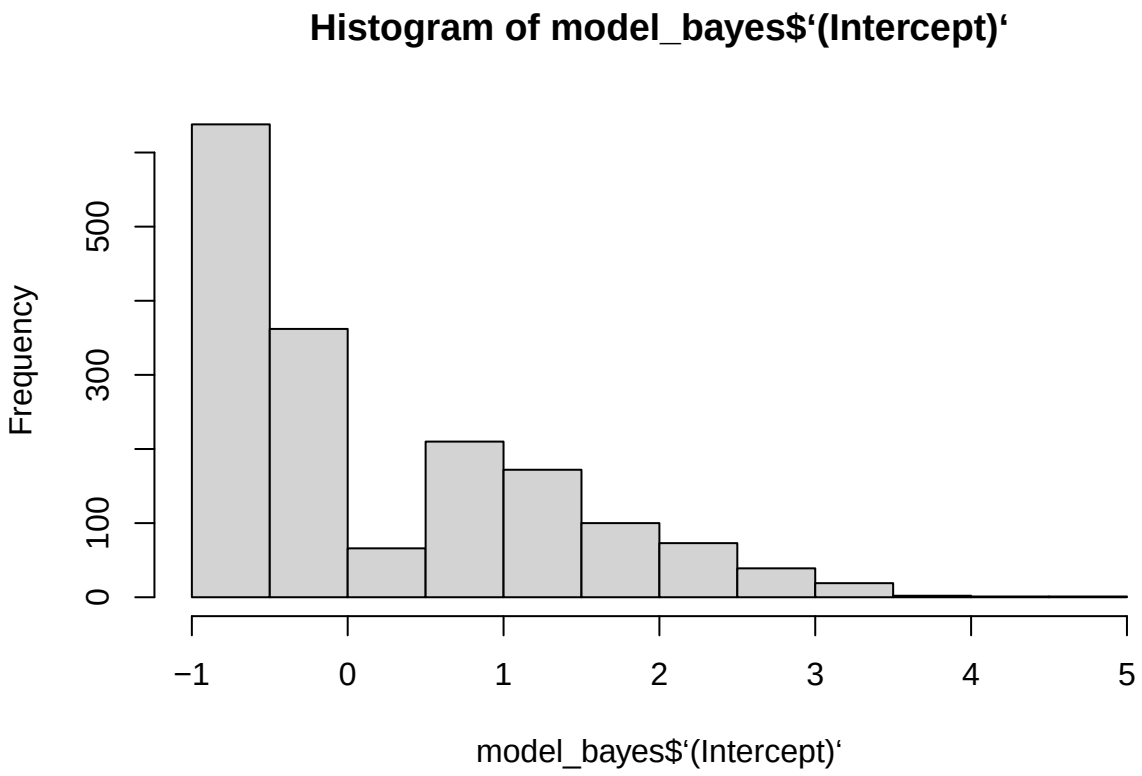
```
## Loading required package: arm
```

```
## Loading required package: MASS
```

```
##
## arm (Version 1.12-2, built: 2021-10-15)
```

```
## Working directory is /Users/gibranerlangga/Documents/Documents - Gibran's MacBook Pro/dartmouth_qbs/
```

```
model_bayes <-REextract(model)
hist(model_bayes$`(Intercept)`)
```



c. Using the same data, fit generalized estimating equation models with a Poisson family and log link and compound symmetry (“exchangeable”) working correlation matrices using the same fixed effects as in 1a.

```
library(geepack)
```

```
## Warning: package 'geepack' was built under R version 4.1.1
```

```
model_gee<-geeglm(Y~Treatment+Year, family=poisson(link='log'),
                  id=ID, corstr="exchangeable", data=skin_data)
summary(model_gee)
```

```
##
## Call:
## geeglm(formula = Y ~ Treatment + Year, family = poisson(link = "log"),
##     data = skin_data, id = ID, corstr = "exchangeable")
##
##  Coefficients:
##             Estimate  Std.err    Wald Pr(>|W|)
## (Intercept) -1.41234  0.10797 171.095   <2e-16 ***
## Treatment    0.14784  0.10943   1.825    0.177
## Year         0.01735  0.02474   0.492    0.483
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Correlation structure = exchangeable
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)    2.647  0.3736
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha   0.3777  0.1111
## Number of clusters:   1683  Maximum cluster size: 5
```

d. Compare the treatment effect estimates with the generalized linear mixed model, and discuss any differences in interpretation.

Estimations in GLMM are used to estimate the individual coefficient separately before getting the average, while estimations in GEE are used to compute the population's average log odds. The Treatment and Year estimates from GLMM are 0.17309 and 0.01827 with standard errors of 0.09978 and 0.01792, while from the GEE we got 0.1478 and 0.0173, with standard errors of 0.1094 and 0.0247. We can see that the estimates we got from GEE are lower than the GLMM, but the the values for standard error are slighly higher compared to the standard errors we got from GLMM.

2. Add the "Exposure" variable to the models above.

a. Evaluate the strength of the association of this variable with the outcome.

```
model_w_exposure <- glmer(Y ~ Treatment + Year + Exposure + (1 | ID), family=poisson(link="log"), data=
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00473363 (tol = 0.002, component 1)
```

```
summary(model_w_exposure)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula: Y ~ Treatment + Year + Exposure + (1 | ID)
##    Data: skin_data
##
##      AIC      BIC   logLik deviance df.resid
##     8098     8132    -4044     8088     7076
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -2.755 -0.351 -0.239 -0.223  6.403
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  ID     (Intercept) 1.42     1.19
## Number of obs: 7081, groups:  ID, 1683
```

```
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.9290     0.1022  -28.67   <2e-16 ***
## Treatment     0.1265     0.0877    1.44     0.15
## Year          0.0197     0.0179    1.10     0.27
## Exposure      0.1979     0.0107   18.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr) Trtmnt Year
## Treatment -0.451
## Year      -0.493  0.011
## Exposure  -0.509 -0.017  0.024
## optimizer (Nelder_Mead) convergence code: 0 (OK)
## Model failed to converge with max|grad| = 0.00473363 (tol = 0.002, component 1)
```

b. Generate missing data indicators for each observation where the missingness probabilities depends on the Exposure variable being above or below a threshold. Apply this to the outcome Y (eg. make it NA), and refit the model. At least 20% of the outcomes should be missing. Compare the missing data rates in each arm. Comment of the difference in treatment effect estimates and confidence intervals for the estimates when applying the ITT analyses to the new dataset with and without using the Exposure variable.

```
library(mice)
```

```
##
## Attaching package: 'mice'

## The following object is masked from 'package:stats':
##
##     filter

## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
# generate missing data, threshold set to 3 for 22% missing data on outcome
skin_data$Y_missing <- ifelse(skin_data$Exposure > 3, NA, skin_data$Exposure)

# try fitting the model on data with missing values

# w exposure
model <- glmer(Y_missing ~ Treatment + Year + Exposure + (1 | ID), family=poisson(link="log"), data=skin
```

```
## boundary (singular) fit: see ?isSingular
```

```
summary(model)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula: Y_missing ~ Treatment + Year + Exposure + (1 | ID)
##    Data: skin_data
##
##      AIC      BIC   logLik deviance df.resid
##    12569    12602    -6279    12559     5483
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -0.1103 -0.0384 -0.0370  0.1308  0.1352
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  ID     (Intercept) 0        0
## Number of obs: 5488, groups:  ID, 1297
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.526318   0.037008  -14.22   <2e-16 ***
## Treatment    0.000293   0.022016    0.01     0.99
## Year         0.000706   0.008302    0.09     0.93
## Exposure     0.561590   0.013577   41.36   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr) Trtmnt Year
## Treatment -0.285
## Year      -0.619  0.015
## Exposure  -0.666 -0.035  0.005
## optimizer (Nelder_Mead) convergence code: 0 (OK)
## boundary (singular) fit: see ?isSingular
```

```
confint(model)
```

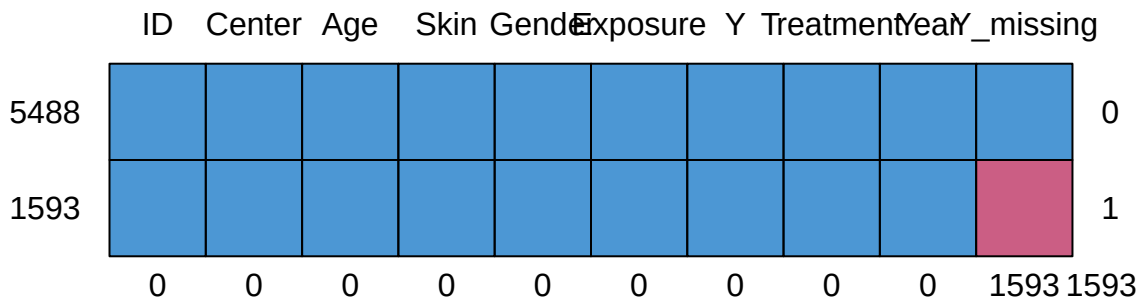
```
## Computing profile confidence intervals ...

##                 2.5 %   97.5 %
## .sig01        0.00000  0.02204
## (Intercept) -0.59904 -0.45397
## Treatment   -0.04286  0.04345
## Year        -0.01558  0.01697
## Exposure     0.53495  0.58817
```

```
# wo exposure
model <- glmer(Y_missing ~ Treatment + Year + (1 | ID), family=poisson(link="log"), data=skin_data)
summary(model)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
```

```
##  Family: poisson  ( log )
## Formula: Y_missing ~ Treatment + Year + (1 | ID)
##    Data: skin_data
##
##      AIC      BIC   logLik deviance df.resid
##    14071    14097    -7031    14063     5484
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -0.369 -0.292 -0.271  0.301  1.140
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  ID     (Intercept) 0.0609   0.247
## Number of obs: 5488, groups:  ID, 1297
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.365011   0.029566   12.35   <2e-16 ***
## Treatment    0.028914   0.026126    1.11     0.27
## Year        -0.000253   0.008353   -0.03     0.98
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr) Trtmnt
## Treatment -0.454
## Year      -0.770  0.013
```

```
#confint(model)

md.pattern(skin_data)
```




```
##      ID Center Age Skin Gender Exposure Y Treatment Year Y_missing
## 5488  1      1   1    1      1        1 1         1    1         1    0
## 1593  1      1   1    1      1        1 1         1    1         0    1
##       0      0   0    0      0        0 0         0    0      1593 1593
```

c. Comment on the need for adjustments for Exposure in this randomized study.

Exposure variable is the only significant variable in the model result above, with coefficient estimate of 0.561.

Data Analysis 2

The dataset reports survival (or censoring) time in years. Individuals with an event (death) are coded as 1, and censoring is coded as 0. The independent variable of primary interest is Treatment (1 = invasive surgery, 0 = less invasive procedure). Other covariates are Female (biological sex is 1 if female, o otherwise), Age, am ordinal Disease Score from 1 to 5 and a Biomarker for which higher is meant to meant worse prognosis.

```
surv_data <- read.delim('ExamSurvData-1.txt')
head(surv_data)
```

```
##    Time Event Treatment Female Age Disease.Score Biomarker
## 1 5.512     0         1      0  22             2       3.1
## 2 4.977     0         0      1  58             2       1.5
## 3 5.466     1         0      1  75             3      14.8
## 4 3.355     0         1      0  31             1       0.5
```

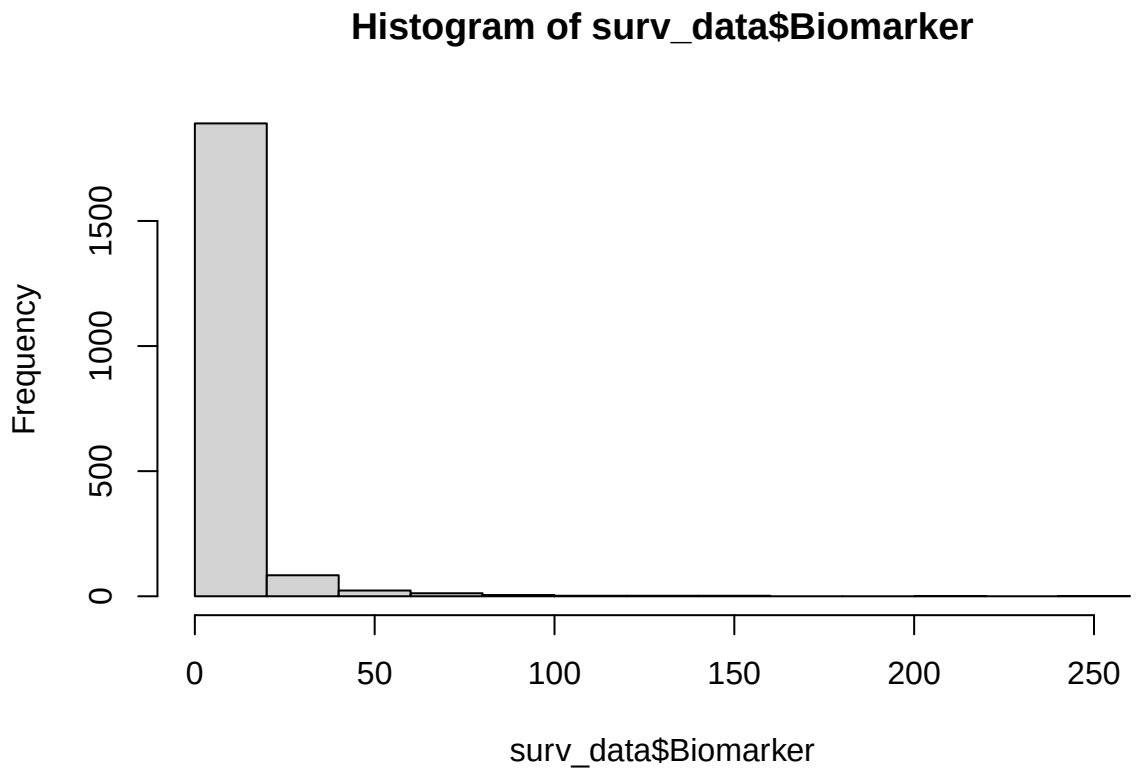
```
## 5 4.603     1         0      1  70             4       6.8
## 6 8.261     0         0      0  63             4       2.4
```

1. a. Describe the distribution of the biomarker.

```
head(surv_data)
```

```
##    Time Event Treatment Female Age Disease.Score Biomarker
## 1 5.512     0         1      0  22             2       3.1
## 2 4.977     0         0      1  58             2       1.5
## 3 5.466     1         0      1  75             3      14.8
## 4 3.355     0         1      0  31             1       0.5
## 5 4.603     1         0      1  70             4       6.8
## 6 8.261     0         0      0  63             4       2.4
```

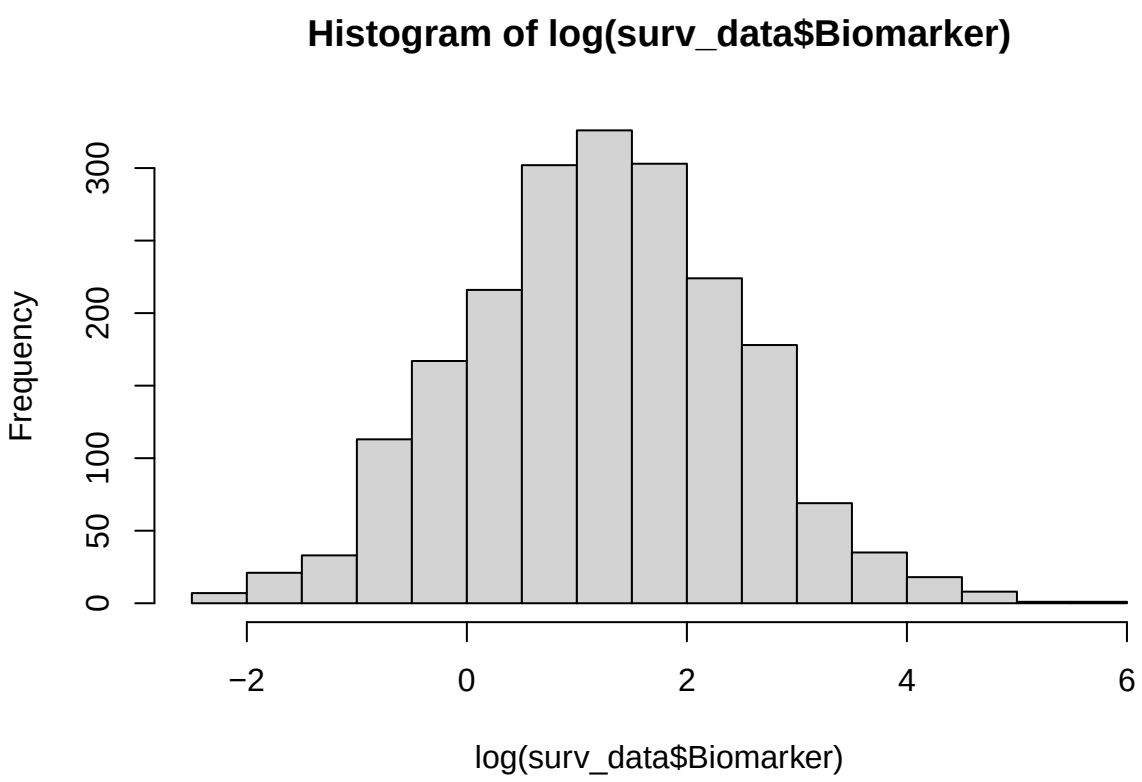
```
hist(surv_data$Biomarker)
```




```
print("The Biomarker distribution seems to be right-skewed. We can apply log-transform on the variable
```

```
## [1] "The Biomarker distribution seems to be right-skewed. We can apply log-transform on the variable
```

```
hist(log(surv_data$Biomarker))
```



b. Report the Pearson and Spearman correlations of the Biomarker with age, disease score and biological sex (or the latter do a two-sample t-test).

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5          v purrr   0.3.4
## v tibble  3.1.5          v dplyr   1.0.7.9000
## v tidyr   1.1.4          v stringr 1.4.0
## v readr   2.0.2          v forcats 0.5.1

## Warning: package 'ggplot2' was built under R version 4.1.1

## Warning: package 'tibble' was built under R version 4.1.1

## Warning: package 'tidyr' was built under R version 4.1.1

## Warning: package 'readr' was built under R version 4.1.1

## Warning: package 'stringr' was built under R version 4.1.1

## Warning: package 'forcats' was built under R version 4.1.1
```

```
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x tidyr::expand() masks Matrix::expand()
## x dplyr::filter() masks mice::filter(), stats::filter()
## x dplyr::lag()    masks stats::lag()
## x tidyr::pack()   masks Matrix::pack()
## x dplyr::select() masks MASS::select()
## x tidyr::unpack() masks Matrix::unpack()
```

```
library(ggpubr)

# age
cor_biomark_age_pearson <- cor.test(surv_data$Biomarker, surv_data$Age,
                                    method="pearson")$estimate
cor_biomark_age_spearman <- cor.test(surv_data$Biomarker, surv_data$Age,
                                     method="spearman")$estimate
```

```
## Warning in cor.test.default(surv_data$Biomarker, surv_data$Age, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
paste("Pearson correlation value between Biomarker and Age is ",
      round(cor_biomark_age_pearson,3))
```

```
## [1] "Pearson correlation value between Biomarker and Age is  0.204"
```

```
paste("Spearman correlation value between Biomarker and Age is ",
      round(cor_biomark_age_spearman,3))
```

```
## [1] "Spearman correlation value between Biomarker and Age is  0.298"
```

```
# disease score
cor_biomark_dis_pearson <- cor.test(surv_data$Biomarker, surv_data$Disease.Score, method="pearson")$esti
cor_biomark_dis_spearman <- cor.test(surv_data$Biomarker, surv_data$Disease.Score, method="spearman")$es
```

```
## Warning in cor.test.default(surv_data$Biomarker, surv_data$Disease.Score, :
## Cannot compute exact p-value with ties
```

```
paste("Pearson correlation value between Biomarker and Disease Score is ",
      round(cor_biomark_dis_pearson,3))
```

```
## [1] "Pearson correlation value between Biomarker and Disease Score is  0.376"
```

```
paste("Spearman correlation value between Biomarker and Disease Score is ",
      round(cor_biomark_dis_spearman,3))
```

```
## [1] "Spearman correlation value between Biomarker and Disease Score is  0.529"
```

```
# sex
cor_biomark_sex_pearson <- cor.test(surv_data$Biomarker, surv_data$Female, method="pearson")$estimate
cor_biomark_sex_spearman <- cor.test(surv_data$Biomarker, surv_data$Female, method="spearman")$estimate
```

```
## Warning in cor.test.default(surv_data$Biomarker, surv_data$Female, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
paste("Pearson correlation value between Biomarker and Sex is ",
      round(cor_biomark_sex_pearson,3))
```

```
## [1] "Pearson correlation value between Biomarker and Sex is  0.003"
```

```
paste("Spearman correlation value between Biomarker and Sex is ",
      round(cor_biomark_sex_spearman,3))
```

```
## [1] "Spearman correlation value between Biomarker and Sex is  0.034"
```

```
# using two-sample t-test
women_biomarker <- surv_data %>%
  filter(Female == 1) %>%
  pull(Biomarker)
men_biomarker <- surv_data %>%
  filter(Female == 0) %>%
  pull(Biomarker)
# Compute t-test
res <- t.test(women_biomarker, men_biomarker)
res
```

```
##
##  Welch Two Sample t-test
##
## data:  women_biomarker and men_biomarker
## t = 0.12, df = 1867, p-value = 0.9
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.161  1.306
## sample estimates:
## mean of x mean of y
##     7.390     7.318
```

c. Develop a model for the association of the biomarker age, disease score and sex. Comment on the findings.

```
model <- lm(Biomarker ~ Age + Disease.Score + Female, data=surv_data)
summary(model)
```

```
##
## Call:
## lm(formula = Biomarker ~ Age + Disease.Score + Female, data = surv_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -16.94  -5.04  -1.16   1.30 231.24
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.7020     1.0580   -1.61     0.11
## Age            -0.0235     0.0240   -0.98     0.33
## Disease.Score   4.1536     0.2712   15.31   <2e-16 ***
## Female         -0.5473     0.5859   -0.93     0.35
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.1 on 2018 degrees of freedom
## Multiple R-squared:  0.142,  Adjusted R-squared:  0.141
## F-statistic:  111 on 3 and 2018 DF,  p-value: <2e-16
```

The only significant independent variable in the model above is Disease Score, with coefficient value of 4.153.

d. Using the model above how does a unit increase in the disease score affect the biomarker.

```
print('One unit increase in disease score variable yields in 4.153 unit increase in Biomarker variable.
```

```
## [1] "One unit increase in disease score variable yields in 4.153 unit increase in Biomarker variable
```

e. For the multivariable model in part c which yields better fit, the log-transformed biomarker or non-transformed.

```
model <- lm(Biomarker ~ Age + Disease.Score + Female, data=surv_data)
summary(model)
```

```
##
## Call:
## lm(formula = Biomarker ~ Age + Disease.Score + Female, data = surv_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -16.94  -5.04  -1.16   1.30 231.24
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.7020     1.0580   -1.61     0.11
## Age            -0.0235     0.0240   -0.98     0.33
## Disease.Score   4.1536     0.2712   15.31   <2e-16 ***
## Female         -0.5473     0.5859   -0.93     0.35
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.1 on 2018 degrees of freedom
## Multiple R-squared:  0.142,  Adjusted R-squared:  0.141
## F-statistic:  111 on 3 and 2018 DF,  p-value: <2e-16
```

```
model <- lm(log(Biomarker) ~ Age + Disease.Score + Female, data=surv_data)
summary(model)
```

```
##
## Call:
## lm(formula = log(Biomarker) ~ Age + Disease.Score + Female, data = surv_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.755 -0.698 -0.001  0.701  3.905
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.05675    0.08297    0.68     0.49
## Age           -0.00202    0.00188   -1.08     0.28
## Disease.Score  0.50373    0.02127   23.68   <2e-16 ***
## Female         0.00706    0.04595    0.15     0.88
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.03 on 2018 degrees of freedom
## Multiple R-squared:  0.288,  Adjusted R-squared:  0.287
## F-statistic:  273 on 3 and 2018 DF,  p-value: <2e-16
```

Log-transformed one performs better, showed by larger adjusted R-squared value compared to the original model.

2. Plot Kaplan-Meier survival curves stratified by Treatment group.

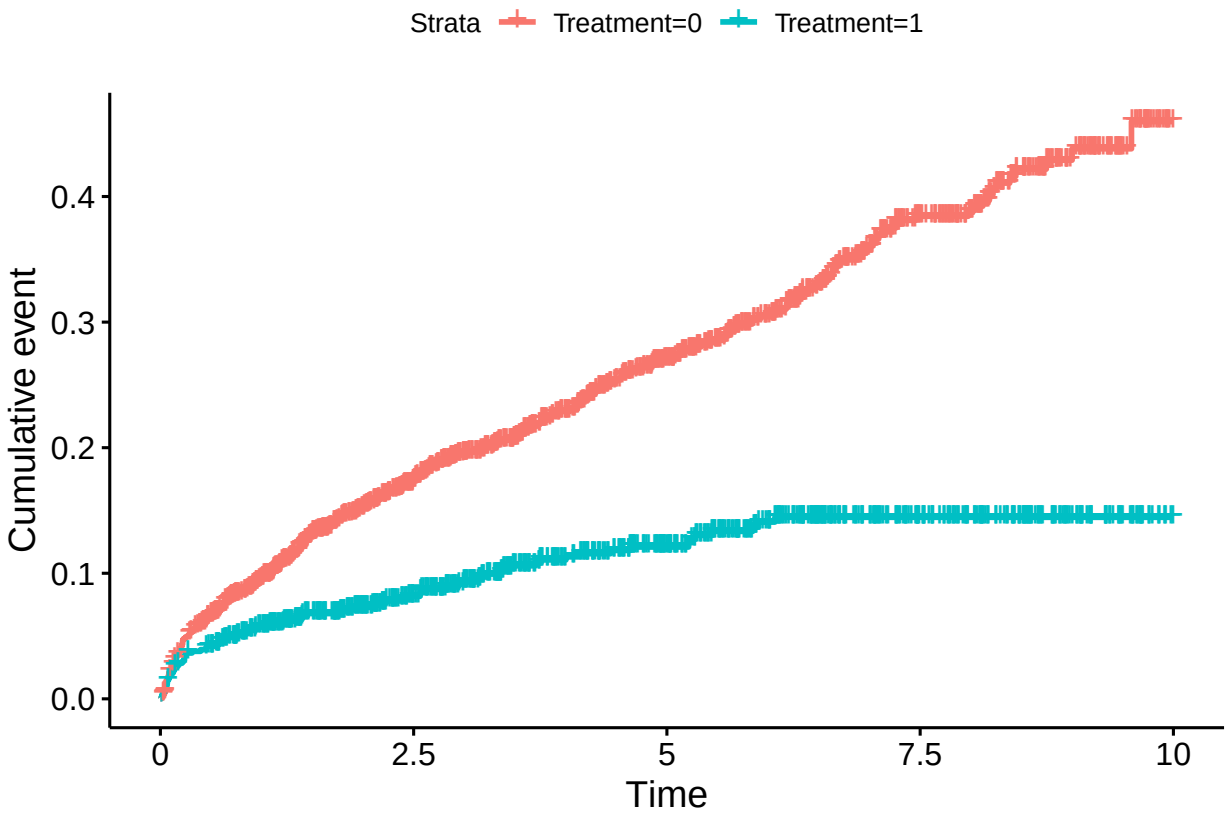
```
library(survival)
library(survminer)
```

```
## Warning: package 'survminer' was built under R version 4.1.1
```

```
##
## Attaching package: 'survminer'
```

```
## The following object is masked from 'package:survival':
##
##     myeloma
```

```
n_death <- surv_data$Event > 0

survival_curve <- survfit(Surv(Time, n_death) ~ Treatment, data = surv_data)
ggsurvplot(survival_curve, fun='event')
```



3. Is there a significant difference in survival between the two treatment groups.

```
wilcox.test(surv_data$Biomarker~surv_data$Female)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  surv_data$Biomarker by surv_data$Female
## W = 490910, p-value = 0.1
## alternative hypothesis: true location shift is not equal to 0
```

```
print('Yes, there is a significant difference between the two treatment groups because the p-value score
```

```
## [1] "Yes, there is a significant difference between the two treatment groups because the p-value sco
```

4. Run a multivariable Cox P.H. model for how the variables Female, Age, Disease Score and Biomarker affect survival.

```
cox_model <- coxph(Surv(Time, n_death) ~ Female + Age + Disease.Score + Biomarker, data = surv_data)
summary(cox_model)
```

```
## Call:
## coxph(formula = Surv(Time, n_death) ~ Female + Age + Disease.Score +
```

```
##     Biomarker, data = surv_data)
##
##   n= 2022, number of events= 425
##
##                    coef exp(coef) se(coef)     z Pr(>|z|)
## Female          0.06292   1.06494  0.09805  0.64     0.52
## Age             0.05224   1.05362  0.00474 11.03  < 2e-16 ***
## Disease.Score   0.17950   1.19662  0.04497  3.99  6.6e-05 ***
## Biomarker      -0.00248   0.99752  0.00302 -0.82     0.41
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##               exp(coef) exp(-coef) lower .95 upper .95
## Female            1.065      0.939     0.879      1.29
## Age               1.054      0.949     1.044      1.06
## Disease.Score     1.197      0.836     1.096      1.31
## Biomarker         0.998      1.002     0.992      1.00
##
## Concordance= 0.723  (se = 0.013 )
## Likelihood ratio test= 299  on 4 df,   p=<2e-16
## Wald test            = 252  on 4 df,   p=<2e-16
## Score (logrank) test = 283  on 4 df,   p=<2e-16
```

```
print("From the model result above, Age and Disease.Score are significant.")
```

```
## [1] "From the model result above, Age and Disease.Score are significant."
```

5. Add Treatment to this model and report the hazard ratio with 95%CI comparing the invasive to less invasive procedure adjusted for the variables in part 4. Is it a statistically significant effect.

```
cox_model_w_treatment <- coxph(Surv(Time, n_death) ~ Female + Age + Disease.Score
                               + Biomarker + Treatment, data = surv_data)
summary(cox_model_w_treatment)
```

```
## Call:
## coxph(formula = Surv(Time, n_death) ~ Female + Age + Disease.Score +
##     Biomarker + Treatment, data = surv_data)
##
##   n= 2022, number of events= 425
##
##                    coef exp(coef) se(coef)     z Pr(>|z|)
## Female          0.06204   1.06401  0.09808  0.63     0.53
## Age             0.05143   1.05277  0.00525  9.80  < 2e-16 ***
## Disease.Score   0.18093   1.19833  0.04514  4.01  6.1e-05 ***
## Biomarker      -0.00247   0.99754  0.00302 -0.82     0.41
## Treatment      -0.04917   0.95202  0.13803 -0.36     0.72
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##               exp(coef) exp(-coef) lower .95 upper .95
## Female            1.064      0.940     0.878      1.29
## Age               1.053      0.950     1.042      1.06
```

```
## Disease.Score     1.198      0.834     1.097      1.31
## Biomarker         0.998      1.002     0.992      1.00
## Treatment         0.952      1.050     0.726      1.25
##
## Concordance= 0.723  (se = 0.013 )
## Likelihood ratio test= 299  on 5 df,   p=<2e-16
## Wald test            = 251  on 5 df,   p=<2e-16
## Score (logrank) test = 283  on 5 df,   p=<2e-16
```

```
exp(cbind(cox_model_w_treatment$coef, confint(cox_model_w_treatment)))
```

```
##                       2.5 % 97.5 %
## Female        1.0640 0.8779  1.290
## Age           1.0528 1.0420  1.064
## Disease.Score 1.1983 1.0969  1.309
## Biomarker     0.9975 0.9916  1.003
## Treatment     0.9520 0.7264  1.248
```

6. Test the proportionality of hazards assumption for each variable in the multivariable model and comment.

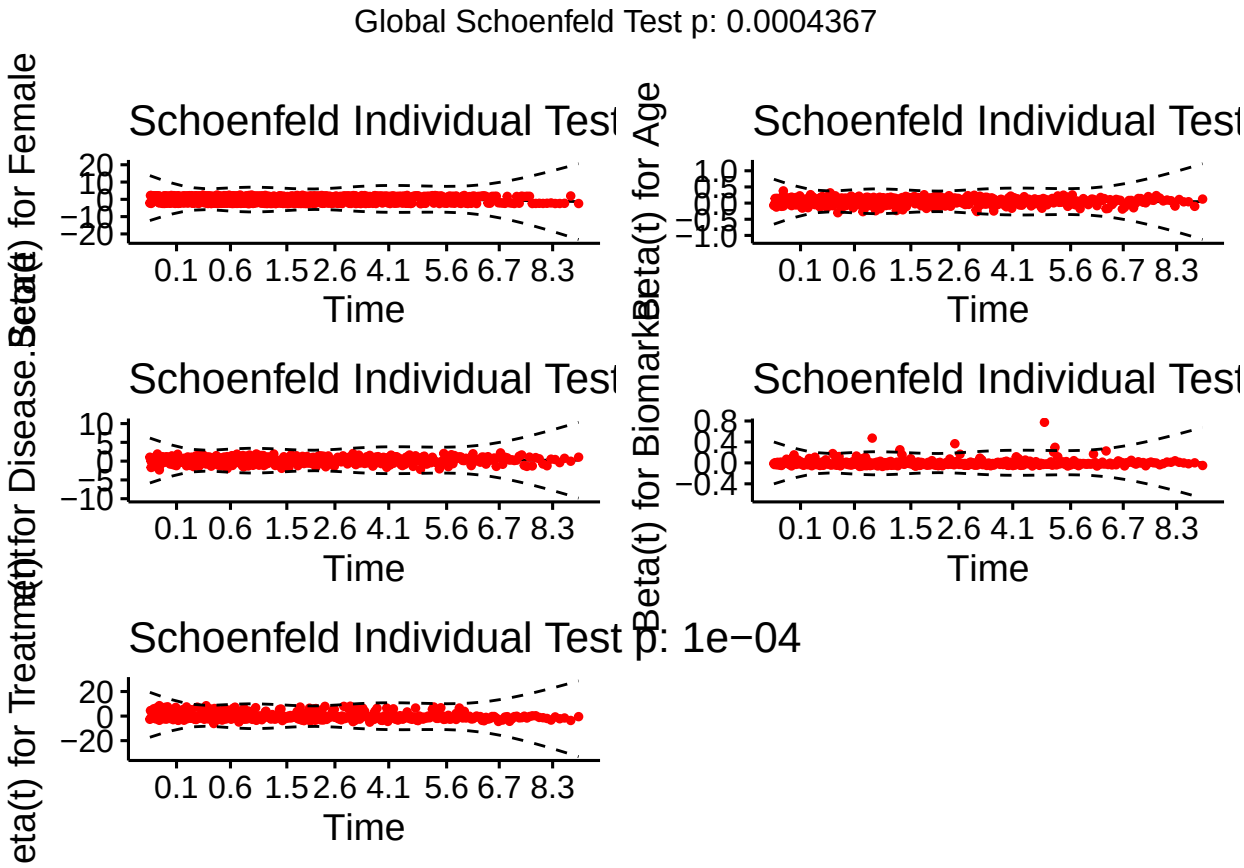
```
library(survminer)

test_prop <- cox.zph(cox_model_w_treatment)
test_prop
```

```
##               chisq df       p
## Female         2.54  1 0.11124
## Age            4.49  1 0.03408
## Disease.Score  4.42  1 0.03544
## Biomarker      3.36  1 0.06669
## Treatment     14.47  1 0.00014
## GLOBAL        22.41  5 0.00044
```

```
print("As seen on the table above, the variables with p-value less than 0.05 are Age, Disease Score, Tr
```

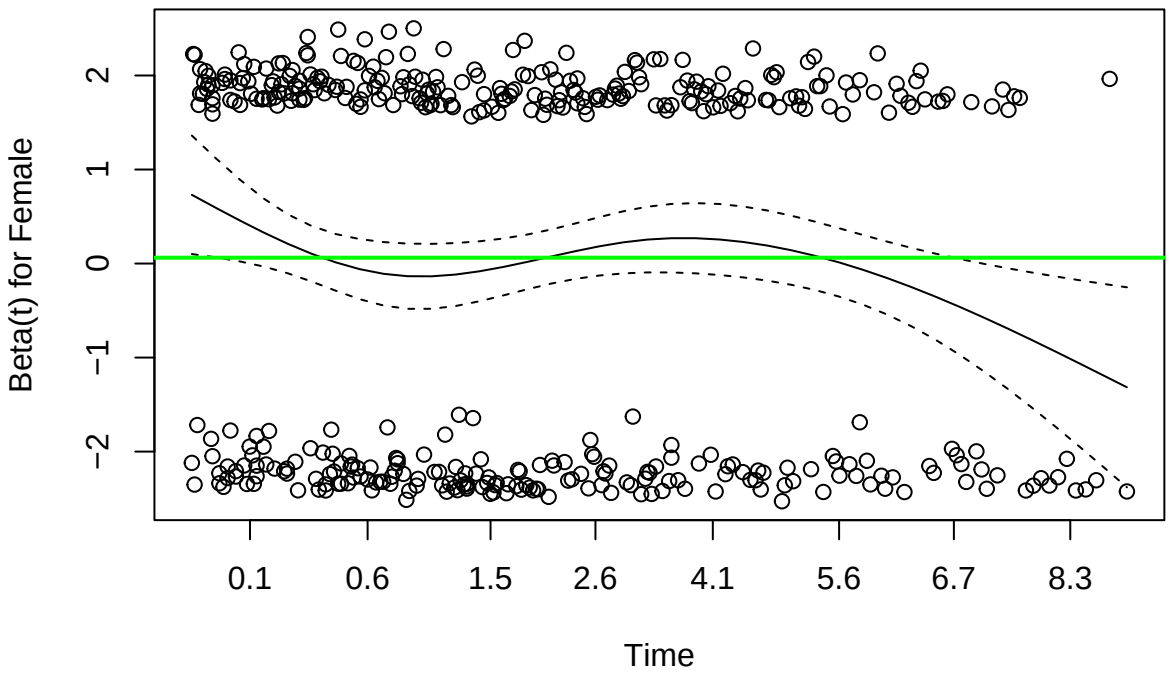
```
## [1] "As seen on the table above, the variables with p-value less than 0.05 are Age, Disease Score, T
```

```
ggcoxzph(test_prop)
```



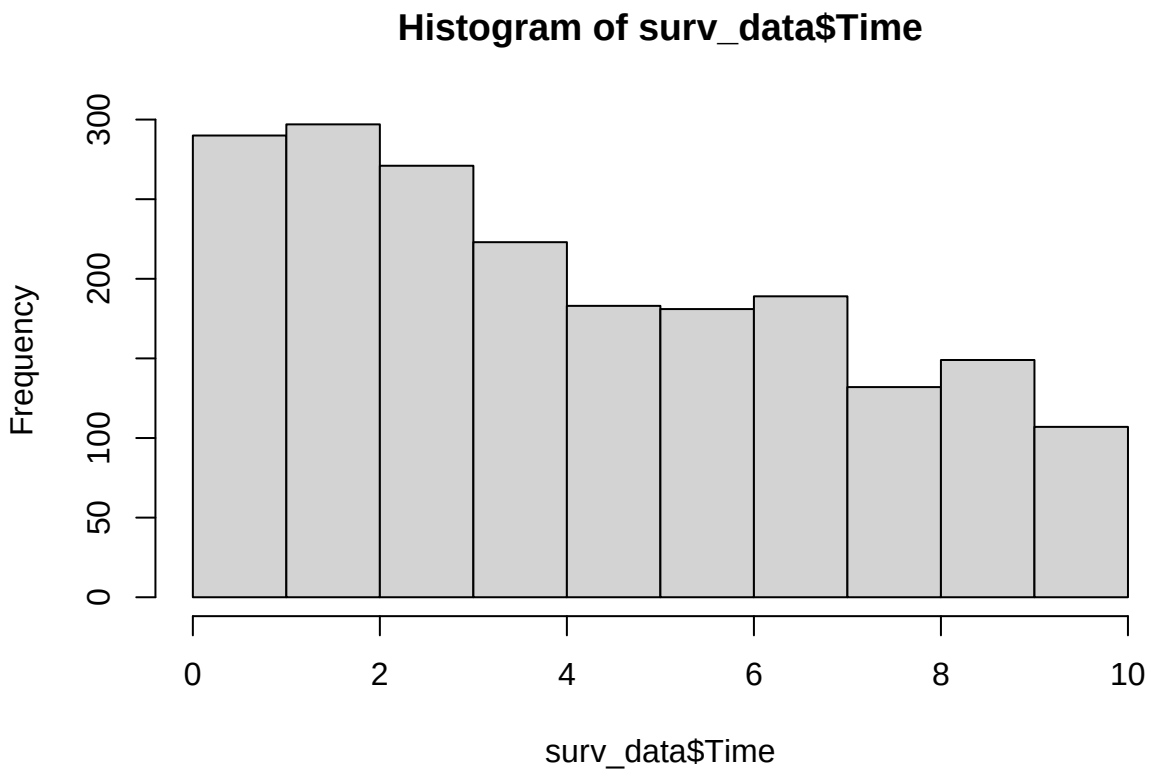
7. Plot the Schoenfeld residuals corresponding to treatment and their smoother as a function of time.

```
plot(test_prop, var=1)
abline(h = coef(cox_model_w_treatment)[1], col = "green", lwd = 2)
```



8. Report the hazard ratio for treatment adjusted for sex, age, disease score and the biomarker for the following time windows, a. < 0.25 years, b. 0.25 to < 1 year and c. 1 year and above.

```
hist(surv_data$Time)
```




```
# less than 0.25
data_1 <- surv_data[surv_data$Time < 0.25,]
n_death_1 <- data_1$Event > 0

# 0.25 to 1 year
data_2 <- surv_data[(surv_data$Time >= 0.25) & (surv_data$Time < 1),]
n_death_2 <- data_2$Event > 0

# more than 1 year
data_3 <- surv_data[surv_data$Time >= 1,]
n_death_3 <- data_3$Event > 0

h_ratio_1 <- coxph(Surv(Time, n_death_1) ~ Treatment + Female + Age +
                     Disease.Score + Biomarker, data = data_1)
# hazard ratio
exp(cbind(h_ratio_1$coef, confint(h_ratio_1)))
```

```
##                       2.5 % 97.5 %
## Treatment     1.1044 0.6344  1.923
## Female        0.7956 0.5073  1.248
## Age           1.0106 0.9846  1.037
## Disease.Score 1.0415 0.8272  1.311
## Biomarker     1.0015 0.9762  1.028
```

```
h_ratio_2 <- coxph(Surv(Time, n_death_2) ~ Treatment + Female + Age +
                     Disease.Score + Biomarker, data = data_2)
exp(cbind(h_ratio_2$coef, confint(h_ratio_2)))
```

```
##                       2.5 % 97.5 %
## Treatment     0.9472 0.5149  1.742
## Female        1.1445 0.7168  1.827
## Age           1.0262 1.0029  1.050
## Disease.Score 1.0721 0.8672  1.325
## Biomarker     0.9981 0.9863  1.010
```

```
h_ratio_3 <- coxph(Surv(Time, n_death_3) ~ Treatment + Female + Age +
                     Disease.Score + Biomarker, data = data_3)
exp(cbind(h_ratio_3$coef, confint(h_ratio_3)))
```

```
##                       2.5 % 97.5 %
## Treatment     0.7074 0.4878  1.026
## Female        0.9563 0.7496  1.220
## Age           1.0554 1.0414  1.070
## Disease.Score 1.2409 1.1092  1.388
## Biomarker     0.9978 0.9906  1.005
```

9. a. Derive a propensity score for Treatment based on sex, age, disease score and biomarker and calculate IWP (inverse weighted propensities).

```
# propensity score
model <- glm(Treatment ~ Female + Age + Disease.Score + Biomarker,
                  family=binomial, data=surv_data)
summary(model)
```

```
##
## Call:
## glm(formula = Treatment ~ Female + Age + Disease.Score + Biomarker,
##     family = binomial, data = surv_data)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.290  -0.756  -0.397   0.783   2.598
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    4.388092   0.234126   18.74   <2e-16 ***
## Female        -0.151400   0.111521   -1.36   0.1746
## Age           -0.106149   0.005519  -19.23   <2e-16 ***
## Disease.Score  0.171656   0.055851    3.07   0.0021 **
## Biomarker      0.000297   0.004613    0.06   0.9487
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 2657.1  on 2021  degrees of freedom
## Residual deviance: 1968.5  on 2017  degrees of freedom
## AIC: 1978
##
## Number of Fisher Scoring iterations: 4
```

```
prop <- model$fit
iwp <- ifelse(surv_data$Treatment, 1/prop, 1/(1-prop))
```

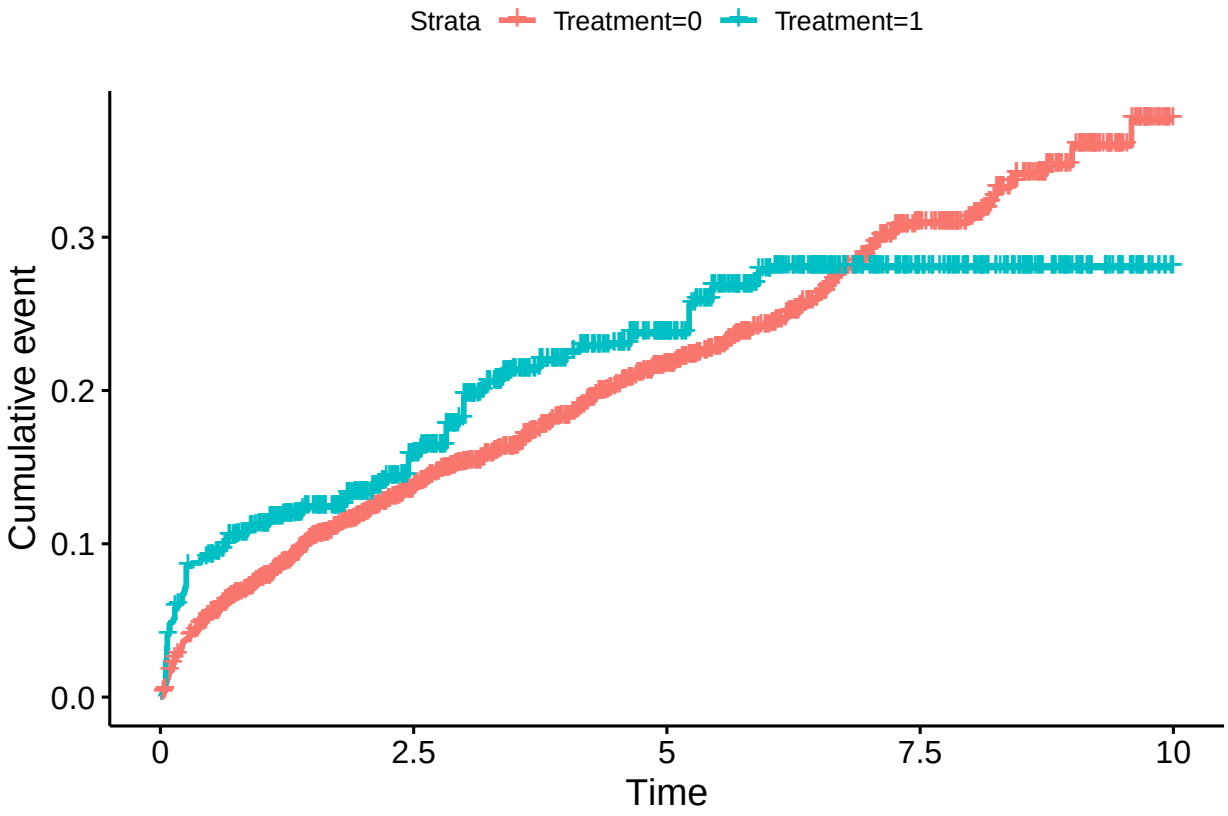
b. What covariates influence treatment selection?

```
print("Age and Disease Score significantly influence Treatment.")
```

```
## [1] "Age and Disease Score significantly influence Treatment."
```

c. Plot Kaplan-Meiers for the two treatment groups weighted by IWP.

```
surv_iwp <- survfit(Surv(Time, n_death) ~ Treatment, data=surv_data, weights=iwp)
ggsurvplot(surv_iwp, fun='event')
```



d. Calculate the hazard ratio for treatment weighted by IWP.

```
h_ratio_iwp <- coxph(Surv(Time, n_death) ~ Treatment, weights = iwp,
                     data=surv_data)
exp(cbind(h_ratio_iwp$coef, confint(h_ratio_iwp)))
```

```
##                   2.5 % 97.5 %
## Treatment 1.062 0.7678  1.469
```

10. Derive the doubly robust estimator of the hazard ratio for treatment by combining a multivariable Cox model with weighting by IWP.

```
h_ratio_all_iwp <- coxph(Surv(Time, n_death) ~ Treatment + Female + Age +
                           Disease.Score + Biomarker, data=surv_data)
exp(cbind(h_ratio_all_iwp$coef, confint(h_ratio_all_iwp)))
```

```
##                        2.5 % 97.5 %
## Treatment     0.9520 0.7264  1.248
## Female        1.0640 0.8779  1.290
## Age           1.0528 1.0420  1.064
## Disease.Score 1.1983 1.0969  1.309
## Biomarker     0.9975 0.9916  1.003
```