# Hw7

Zhuodiao Kuang

2023-11-17

## Problem 1

Control the age at transplant for those subjects who received the transplant in the heart transplant data set.

The age should be the age at baseline before transplant. The age at the time of transplant should be calculated using baseline_age+wait_time. **Clarification:** The age variable in the heart transplant dataset is the age when the subject enrolled to the study. The risk of death is likely to be associated with age. The treatment benefit of heart transplant, if any, may also depend upon the age at transplant.

Problem 1 in Homework 7 asks to understand the transplant's effect when controlling the age at transplant in the time-varying covariate model. In the Cox model, age should be the age at enrollment of the study before transplant. **If subjects received transplants, the age should be changed to the age at the time of transplant.**)

## Solution 1

### Import and clean the data.

To obtain the wait time for every patient, I had to manipulate the data in this case. Upon observing that the age was centered around 48, I restored it to its initial scale. I use the original scale of the baseline age plus the wait time divided by 365.25 to determine the age at transplant for those who received transplantation; for those who did not receive transplantation, the age at transplant is equal to the baseline age.

```
heart_dat = readxl::read_excel("Datasets.xlsx", sheet = "Heartdata")
heart_dat$WAITING_TIME_FOR_TRANSPLANT =
  ifelse(is.na(
    heart_dat$WAITING_TIME_FOR_TRANSPLANT), 0,
    heart_dat$WAITING_TIME_FOR_TRANSPLANT)
heart_dat$AGE_T = heart_dat$AGE + ((heart_dat$WAITING_TIME_FOR_TRANSPLANT)/365.25)

heart_dat$TRANSPLANT_STATUS = as.factor(heart_dat$TRANSPLANT_STATUS)
heart_dat$SURVIVAL_STATUS = as.numeric(heart_dat$SURVIVAL_STATUS)
heart_dat$SURVIVAL_TIME = as.numeric(heart_dat$SURVIVAL_TIME)
heart_dat$PATIENT_ID = as.integer(heart_dat$PATIENT_ID)
```

Two covariates, the indicator of transplantation and the time-varying variable `AGE_T` are included when building the cox model:

```
heart.fit = coxph(Surv(SURVIVAL_TIME, SURVIVAL_STATUS==1) ~
                    TRANSPLANT_STATUS + AGE_T, data = heart_dat)
heart.sum = summary(heart.fit)
rownames(heart.sum$coefficients) = c("Transplant Status = 1", "Age(t)")
heart.sum$coefficients |>
  kable("latex",
        digits = 4,
        escape = F,
        booktabs = T,
        caption = "Regresion Coefficients Estimates of the Time-Varying Cox Model") |>
  kable_styling(position = "center", latex_options = "hold_position")
```

Table 1: Regresion Coefficients Estimates of the Time-Varying Cox Model

| | coef | exp(coef) | se(coef) | z | Pr(>\|z\|) |
|---|---|---|---|---|---|
| Transplant Status = 1 | -1.7947 | 0.1662 | 0.2717 | -6.6048 | 0e+00 |
| Age(t) | 0.0599 | 1.0617 | 0.0153 | 3.9223 | 1e-04 |

From the regression coefficients result above, we can see that when controlling the age at transplant, the transplant's effect shows significant, and it significantly lower the risk of event, the risk reduction is about 83.4%.

# Problem 2

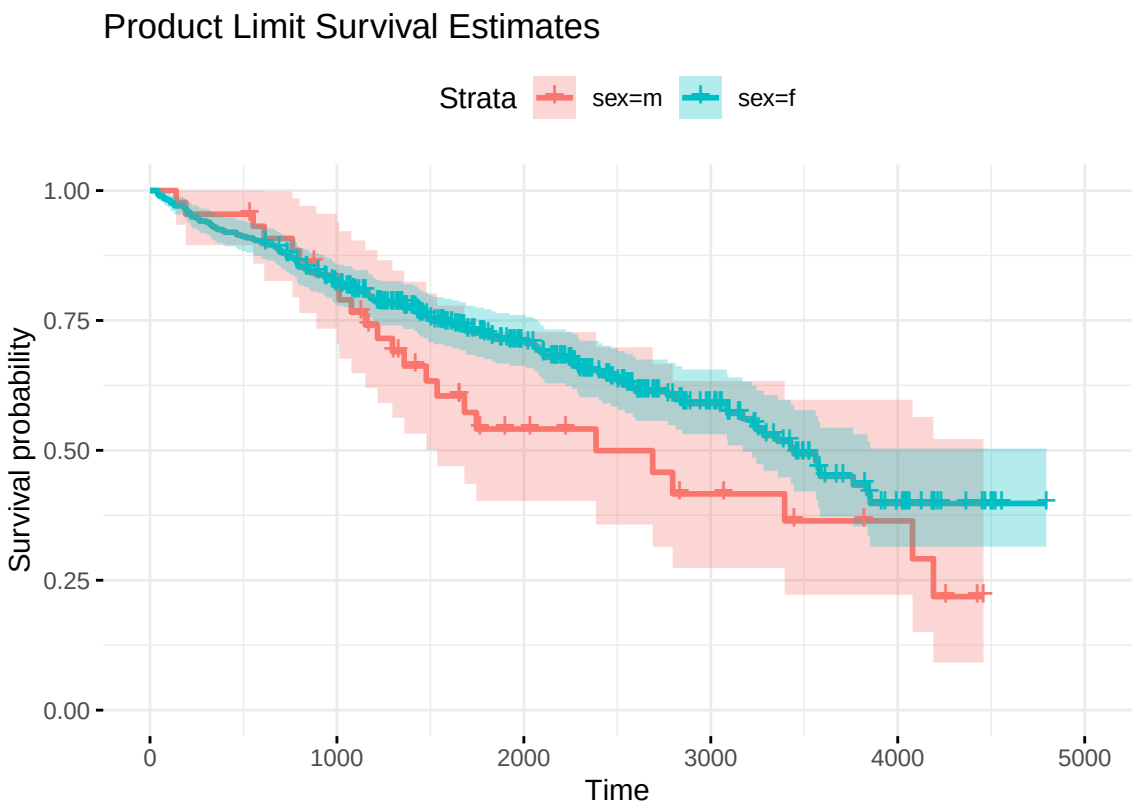
Check proportional hazard assumption between the two sex groups in the PBC dataset. Provide the results of the checking.

### a.Using log{-logS(t,Z)}

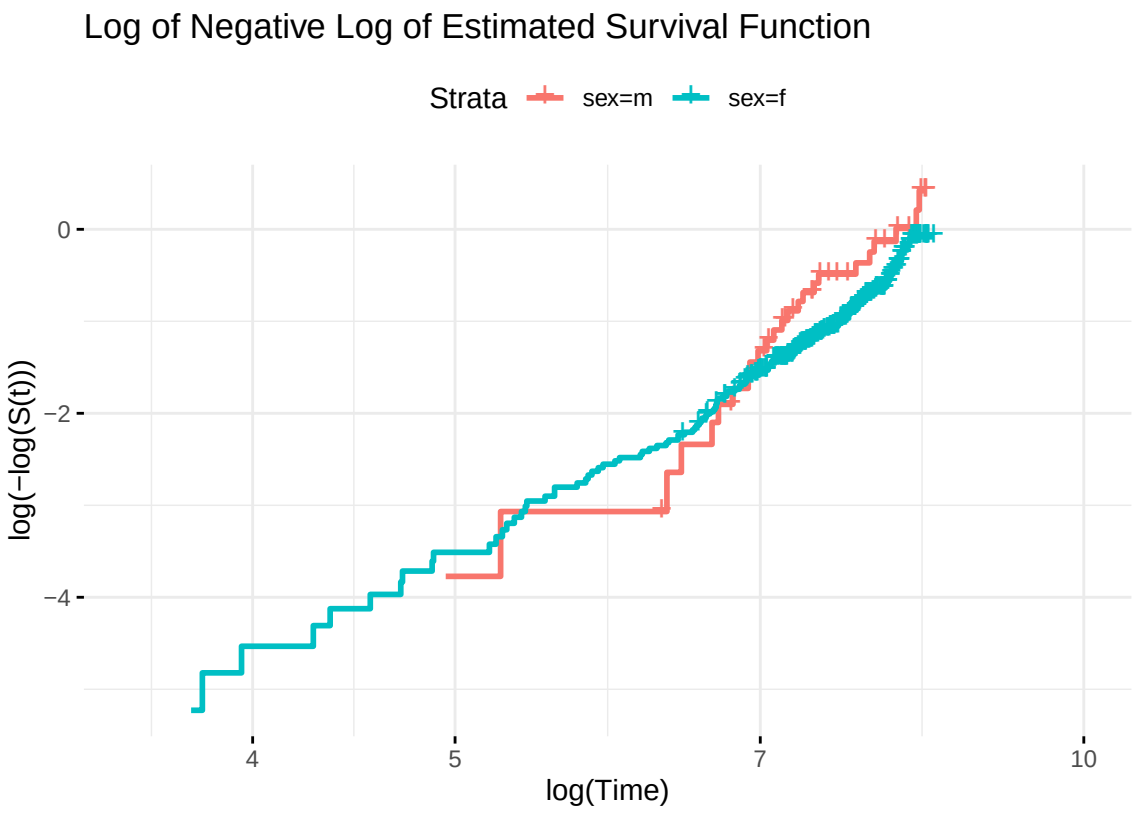
The result of the checking of proportional hazard assumption using $\log\{-\log S(t, Z)\}$ is shown in the following plot:

```
library(ggsurvfit)
library(survminer)
pbc_survfit = survfit(Surv(time, status==2)~ sex, data = pbc)
pbc_survfit_log = survfit(Surv(log(time+1), status==2)~ sex, data = pbc)

splots <- list()
splots[[1]] <- ggsurvplot(pbc_survfit,
                          data = pbc,
                          risk.table = FALSE,
                          ggtheme = theme_minimal(),
                          conf.int = T)
splots[[2]] <- ggsurvplot(pbc_survfit_log,
                          data = pbc,
                          fun = "cloglog",
                          risk.table = FALSE,
                          xlab = "log(Time)",
                          ggtheme = theme_minimal(),
```

```
                          xlim = c(3.5,10))

splots[[1]]$plot + labs(title = "Product Limit Survival Estimates")
```




```
splots[[2]]$plot + labs(title = "Log of Negative Log of Estimated Survival Function")
```

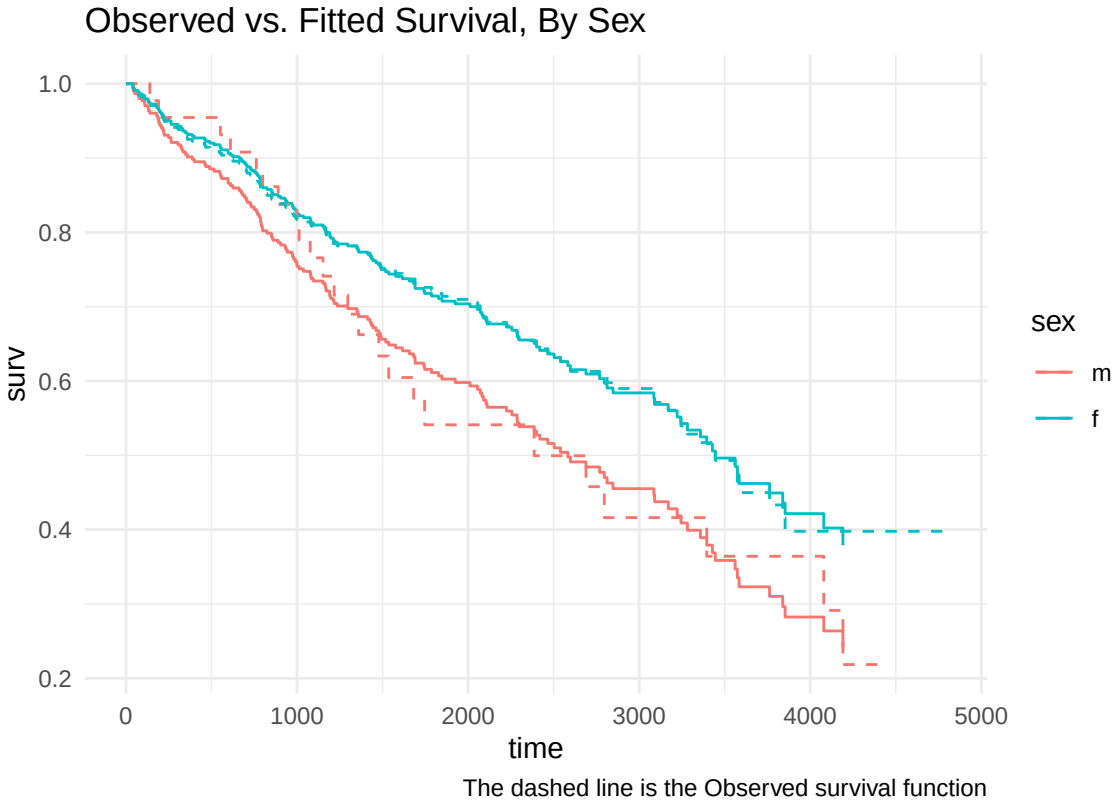



From the plot above, we can found that the $\log\{-\log S(t, Z)\}$ are not two parallel lines, there exists a cross-over at about time = 1000. That means the hazard ratio between male and female is not proportional.

**b.Plot the observed and fitted**

```
obsfit_plots = list()
obsfit_plots[[1]] = ggsurvplot(pbc_survfit,
                               data = pbc,
                               risk.table = FALSE,
                               ggtheme = theme_minimal())
obsfit_plots[[2]] = ggadjustedcurves(coxph(Surv(time, status==2)~ sex,
                                           data = pbc),
                                     variable = "sex",
                                     data = pbc,
                                     ggtheme = theme_minimal())

cox_fit_surv_dat = obsfit_plots[[2]]$data
cox_fit_surv_dat$sex = cox_fit_surv_dat$variable

km_fit_surv_dat  = obsfit_plots[[1]]$plot$data


p <- ggplot(cox_fit_surv_dat, aes(x = time,
                                  y = surv,
                                  group = sex,
```

```
                                    color = sex)) + geom_step()
p + geom_step(data = km_fit_surv_dat, aes(x = time,
                                          y = surv,
                                          group = sex,
                                          color = sex),
              lty = 2) +
  labs(title = "Observed vs. Fitted Survival, By Sex",
  caption = "The dashed line is the Observed survival function") +
  theme_minimal()
```




From the plot above, we can see that the fitted survival function of female group is pretty close to the observed survival function. However, the fitted survival function of the male group cross the observed survival function. That means the proportional assumption does not hold.

### c.Including a couple of continuous variables and check the interaction with time

In order to investigate the interaction of the continuous variables with time, I included additional variables `albumin`, `bili`, `ast`, `copper` and `protime` and their interaction with time (in log scale) into the proportional hazard model. I choose those parameter based on the stepwise results from the last homework.

```
pbc_interaction_fit =
  coxph(Surv(time, status==2)~
          sex + albumin + bili + ast +
          copper + protime +
          log(time):albumin + log(time):bili +
```

```
          log(time):ast + log(time):copper +
          log(time):protime,
          data = pbc)
summary(pbc_interaction_fit)$coefficients|>
  kable("latex",
        digits = 4,
        escape = F,
        booktabs = T,
caption = "Regresion Coefficients Estimates of the Cox Model with Time Interactions") |>
  kable_styling(position = "center",
                latex_options = "hold_position")
```

Table 2: Regresion Coefficients Estimates of the Cox Model with Time Interactions

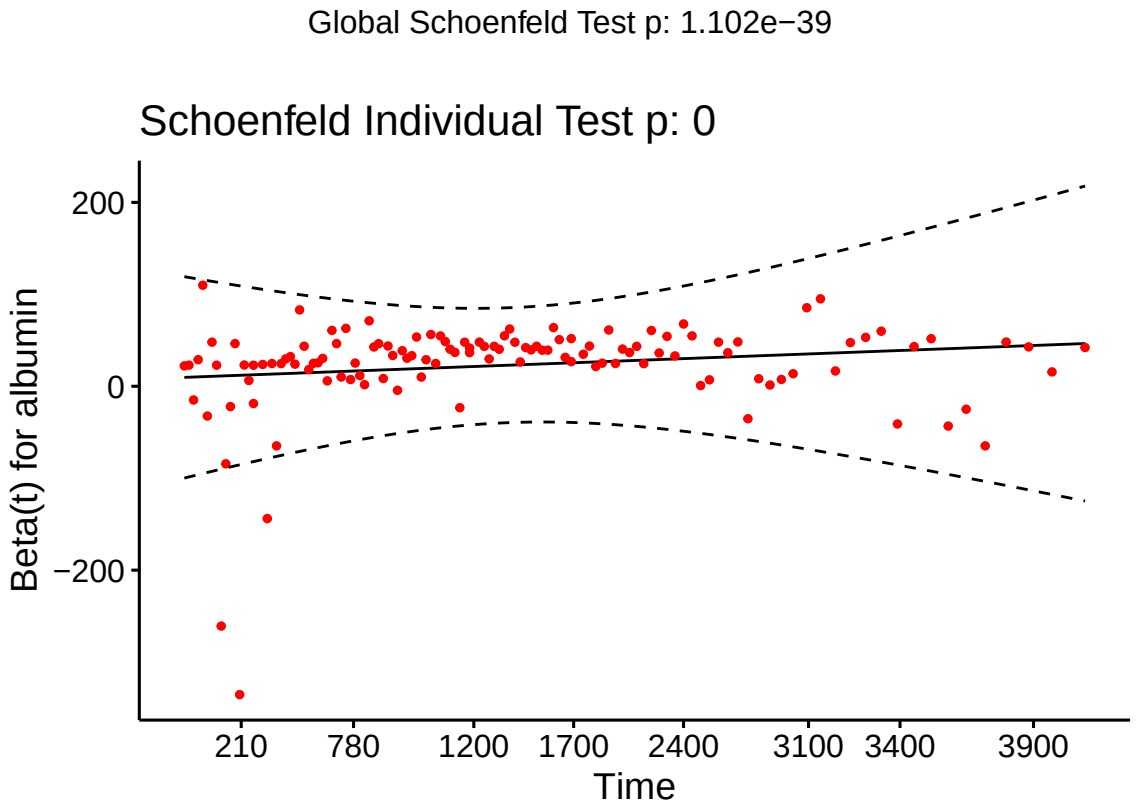
| | coef | exp(coef) | se(coef) | z | Pr(>\|z\|) |
|---|---|---|---|---|---|
| sexf | -0.5014 | 6.057000e-01 | 0.3276 | -1.5306 | 0.1259 |
| albumin | 22.9591 | 9.354306e+09 | 2.8063 | 8.1813 | 0.0000 |
| bili | -1.2217 | 2.947000e-01 | 0.2493 | -4.9013 | 0.0000 |
| ast | 0.0241 | 1.024400e+00 | 0.0208 | 1.1580 | 0.2469 |
| copper | 0.0315 | 1.032000e+00 | 0.0127 | 2.4813 | 0.0131 |
| protime | 10.3510 | 3.128905e+04 | 1.1481 | 9.0158 | 0.0000 |
| albumin:log(time) | -3.2599 | 3.840000e-02 | 0.3893 | -8.3737 | 0.0000 |
| bili:log(time) | 0.1801 | 1.197300e+00 | 0.0362 | 4.9787 | 0.0000 |
| ast:log(time) | -0.0032 | 9.968000e-01 | 0.0028 | -1.1576 | 0.2470 |
| copper:log(time) | -0.0040 | 9.960000e-01 | 0.0018 | -2.2024 | 0.0276 |
| protime:log(time) | -1.3975 | 2.472000e-01 | 0.1599 | -8.7426 | 0.0000 |

From the regression summary above, we can see that except `ast`, all the selected variable has a significant effect and interaction with time.

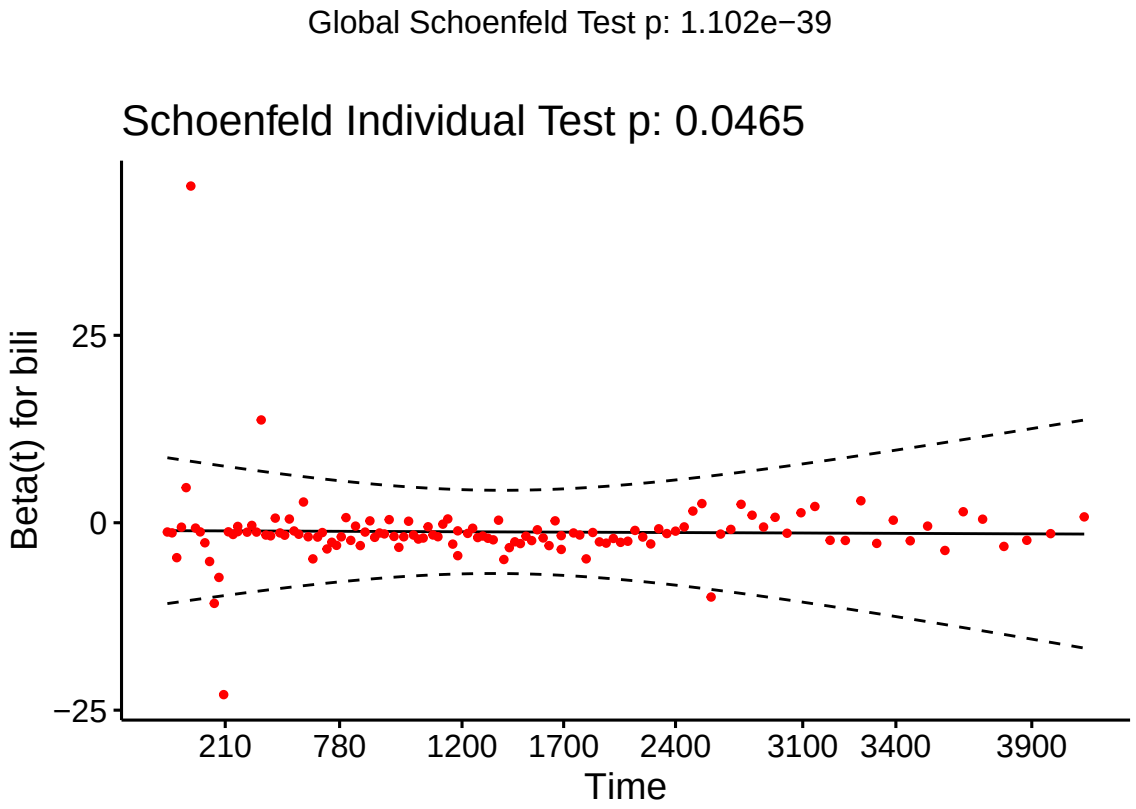
**d.Plot the Schoenfeld residual of the fitted model with two continuous covariates in problem 2.c.**

I plot the Schoenfeld residual of the fitted model with two continuous covariates, `albumin` and `copper`, below.

```
ggcoxzph(cox.zph(pbc_interaction_fit),
         var = c("albumin"),df = 2, nsmo = 1000)
```




```
ggcoxzph(cox.zph(pbc_interaction_fit),
         var = c("bili"),df = 2, nsmo = 1000)
```




From the above plot and the Schoenfeld individual test p-value, we can see that both the residual plots has a non-zero slope regression line, and the p-value of both tests are less than 0.05, which means both the covariates do not meet the PH assumption.