Survival Analysis with Applications in Medicine: Take-home examination

2025-03-28

A. Weibull regression models

Q1. For a proportional hazards model with $S_0(t)$ that is a Weibull distribution, show that survival $S(t)$ is also from a Weibull distribution.

Given a proportional hazards model with survival function $S(t|x) = S_0(t)^{\exp(\beta x)}$

for time t and a given covariate x, $S_0(t)$ is the baseline survival function, and β is the log hazard ratio.

We have the Weibull survival function $S_0(t) = \exp(-\lambda t^k)$ for a scale parameter λ and a shape parameter k.

Substituting $S_0(t)$ in the proportional hazards model, we get the survival function as

$S(t|x) = \left[\exp(-\lambda t^k)\right]^{\exp(\beta x)}$

$\Rightarrow S(t|x) = \exp(-\lambda t^k \exp(\beta x))$

which is in form $S(t|x) = \exp(-\tilde{\lambda}_a t^k)$ where $\tilde{\lambda}_a = \lambda \exp(\beta x)$ is the new scale parameter for the same shape parameter k.

Hence, the survival function $S(t|x)$ for a proportional hazards model is also from a Weibull distribution.

Q2. For an accelerated failure time model with $S_0(t)$ that is a Weibull distribution, show that survival $S(t)$ is also from a Weibull distribution.

Given an accelerated failure time model with survival function $S(t|x) = S_0(t \exp(-\tilde{\beta} x))$

for time t and a given covariate x, $S_0(t)$ is the baseline survival function, and $\tilde{\beta}$ is the log time ratio.

Substituting $S_0(t)$ in the accelerated failure time model, we get the survival function as

$S(t|x) = \exp(-\lambda \left(t \exp(-\tilde{\beta} x)\right)^k)$

$\Rightarrow S(t|x) = \exp(-\lambda t^k \exp(-k\tilde{\beta} x))$

which is in form $S(t|x) = \exp(-\tilde{\lambda}_b t^k)$ where $\tilde{\lambda}_b = \lambda \exp(-k\tilde{\beta} x)$ is the new scale parameter for the same shape parameter k.

Hence, the survival function $S(t|x)$ for an accelerated failure time model is also from a Weibull distribution.

Q3. What is the relationship between β and $\tilde{\beta}$ if both models have a Weibull baseline survival function?

For the proportional hazards model, we have $S(t|x) = \exp(-\lambda t^k \exp(\beta x))$.

For the accelerated failure time model, we have $S(t|x) = \exp(-\lambda t^k \exp(-k\tilde{\beta} x))$.

Equating the two models, we get $\exp(\beta x) = \exp(-k\tilde{\beta} x)$.

Taking the natural logarithm of both sides, we get $\beta x = -k\tilde{\beta} x$.

Therefore, the relationship between β and $\tilde{\beta}$ is $\beta = -k\tilde{\beta}$ if both models have a Weibull baseline survival function.

B: Interval-censored likelihood

For a data tuple (t_i, u_i, v_i) where t_i is the (left truncated) delayed entry time, and the event is observed in the interval $(u_i, v_i]$ for an individual i.

Q1a. Express the log-likelihood in terms of Survival function $S(t)$ at time t:

The Likelihood for the interval-censored data: $u_i < T \leq v_i$ for an entry time t_i is given by:

$L_i = \frac{S(u_i) - S(v_i)}{S(t_i)}$

where $S(t)$ is the Survival function.

This represents the probability that the event occurs in the interval $(u_i, v_i]$ given that the event has not occurred before u_i and the individual survives up to t_i and is at risk at time t_i.

Hence, the log-likelihood in terms of Survival function $S(t)$ at time t for this data tuple is given by $\log(S(u_i) - S(v_i)) - \log(S(t_i))$.

Q1b. Express the log-likelihood in terms of the hazard function $h(t)$ at time t:

Now, for the derivation of the log-likelihood in terms of the hazard function $h(t)$ at time t, we need to express the Survival function $S(t)$ in terms of the hazard function $h(t)$.

We know that, the survival function $S(t)$ is given by $-log(S(t)) = H(t)$,

where $H(t)$ is the cumulative hazard function, and $H(t) = \int_0^t h(u)du$ for the hazard function $h(t)$.

Therefore, the log-likelihood in terms of the hazard function $h(t)$ at time t is:

$\log L_i = \log(S(u_i) - S(v_i)) - \log(S(t_i))$

$\Rightarrow \log L_i = \log(exp(-H(u_i)) - exp(-H(v_i))) - \log(exp(-H(t_i)))$

$\Rightarrow \log L_i = \log(exp(-H(v_i)) * (exp(H(v_i) - H(u_i)) - 1)) + H(t_i)$

$\Rightarrow \log L_i = \log(exp(-H(v_i))) + \log(exp(H(v_i) - H(u_i)) - 1) + H(t_i)$

$\Rightarrow \log L_i = H(t_i) - H(v_i) + \log(exp(H(v_i) - H(u_i)) - 1)$

$\Rightarrow \log L_i = \int_0^{t_i} h(t)dt - \int_0^{v_i} h(t)dt + \log(exp(\int_{u_i}^{v_i} h(t)dt) - 1)$

$\Rightarrow \log L_i = -\int_{t_i}^{v_i} h(t)dt + \log(exp(\int_{u_i}^{v_i} h(t)dt) - 1)$

Hence, the log-likelihood in terms of the hazard function $h(t)$ at time t is given by $-\int_{t_i}^{v_i} h(t)dt + \log(exp(\int_{u_i}^{v_i} h(t)dt) - 1)$.

Q2. Can you express these data using the Surv function from the survival package? If so, show an example; if not, explain why.

The `Surv` function is used to create a survival object that represents the survival time of an individual. It takes the form `Surv(time, event)` where `time` is the survival time and `event` is the event indicator.

For interval-censored data, we can use the `Surv` function as `Surv(time, time2, type = "interval2")` where `time` is the start of the interval, `time2` is the end of the interval and type `interval2` is used to indicate interval-censored data effectively.

Although, the `Surv` function doesn't support left truncation directly, we can filter out the left truncated data by taking the maximum of the entry time and the left truncation time.

In conclusion, we cannot directly use the `Surv` function to directly express left truncated interval censored data, but we can indirectly use it to express by filtering beforehand.

Here is an example of how to express left-truncated interval-censored data indirectly using the `Surv` function:

```
library(survival)

# sample data for the given data tuple structure (t_i, u_i, v_i)
d = data.frame(left_truncation_time = c(1,1,3,3,3),
               entry_time = c(0,0,2,2,2),
               exit_time = 1:5,
               event = c(1,0,1,0,1))

# filter out left truncated data
d$entry_time = pmax(d$entry_time, d$left_truncation_time)

# interval-censored data
with(d, Surv(entry_time, exit_time, type="interval2"))
```

```
## [1] 1      [1, 2] 3      [3, 4] [3, 5]
```

C: Truncated distributions

For a continuous random variable T, we have the survival function $S(t) = P(T > t)$.

Let the quantile function $Q(p)$ such that $Q(p) = S^{-1}(1-p) \Rightarrow 1 - p = S(Q(p)) = P(T > Q(p))$.

Assume that we have a truncated distribution with survival function $P(T > t|T > t_0)$ for left truncation time t_0.

Q1. Express the survival function for the truncated distribution in terms of the survival function for T.

The survival function for the truncated distribution is given by:

$P(T > t|T > t_0)$

$\Rightarrow \frac{P(T>t, T>t_0)}{P(T>t_0)}$ by definition of conditional probability.

$\Rightarrow \frac{P(T>t)}{P(T>t_0)}$ because $(T > t, T > t_0)$ implies $(T > t)$ as t_0 is the left truncation time.

Since, $P(T > t) = S(t)$ and $P(T > t_0) = S(t_0)$ by the survival function for T,

the survival function for the truncated distribution is given by $\frac{S(t)}{S(t_0)}$.

Q2. For the truncated distribution, what is the quantile function $Q(p|t_0)$ that solves $P(T > t|T > t_0) = 1 - p$ for t in terms of the survival and quantile functions for T at quantile (probability) p?

Given $P(T > t|T > t_0) = 1 - p$,

we know that $P(T > t|T > t_0) = \frac{S(t)}{S(t_0)} = 1 - p$.

Therefore, the quantile function $Q(p|t_0) = t$ that solves $P(T > t|T > t_0) = 1 - p$:

$S(t) = (1 - p)S(t_0)$

$\Rightarrow t = S^{-1}((1-p)S(t_0))$

$\Rightarrow Q(p|t_0) = S^{-1}((1-p)S(t_0))$.

To solve for $S^{-1}((1-p)S(t_0))$, we know that $Q(p) = S^{-1}(1-p) \rightarrow Q(1-p) = S^{-1}(p)$.

$\Rightarrow Q(p|t_0) = S^{-1}((1-p)S(t_0)) = Q(1-(1-p)S(t_0))$.

i.e., the p-th quantile of left truncated distribution is the $(1-(1-p)S(t_0))$-th quantile of the original distribution.

Q3. Using this algorithm, write, run and report on R code to calculate the 0.4 quantile from a truncated log-normal distribution where $T \sim LogNormal(\mu = 1, \sigma^2 = 1.2^2)$ for a log-normal distribution with mean μ and standard deviation σ on the log scale for $t_0 = 2$.

We can derive the quantile function $Q(p = 0.4|t_0 = 2)$ as follows:

$Q(p|t_0) = Q(1-(1-p)S(t_0))$.

To derive $S(t_0)$, we know that $S(t_0) = P(T > t_0) = 1 - P(T \leq t_0) = 1 - F(t_0)$,

where $F(t)$ is the cumulative distribution function for T.

For a log-normal distribution of T, we can compute $S(t_0)$ using `plnorm` function in R.

Then, we can substitute the $S(t_0)$ value back to $Q(1-(1-p)S(t_0))$ and compute the quantile using `qlnorm` function in R for the probability $p = 0.4$.

```
#' @param p is the probability
#' @param meanlog mean on the log scale
#' @param sdlog standard deviation on the log scale
#' #' @param t0 left truncation time
#' @return the quantile from a truncated log-normal distribution

f = function(p, meanlog, sdlog, t0) {
  S2 = 1 - plnorm(t0, meanlog, sdlog)
  Q = qlnorm(1 - (1-p)*S2, meanlog, sdlog)
  return(Q)
}

# function call
f(p = 0.4, meanlog = 1, sdlog = 1.2, t0 = 2)
```

```
## [1] 4.171994
```

Hence, the 0.4 quantile from a truncated log-normal distribution with $T \sim LogNormal(\mu = 1, \sigma^2 = 1.2^2)$ for $t_0 = 2$ is approximately 4.172.

Q4. Check your value of the 0.4 quantile from the truncated log-normal distribution in question C3 by given random sampling code.

Given `R` code to return a vector of random numbers t sampled from a truncated log-normal distribution:

```
#' @param n the number of random numbers
#' @param meanlog mean on the log scale
#' @param sdlog sd on the log scale
#' @param t0 left truncation time(s)
#' @return vector of random numbers drawn from a truncated log-normal distribution
rtrunc_lnorm = function(n, meanlog, sdlog, t0) {
    y = rlnorm(n, meanlog, sdlog)
    while (any(y<t0))
```

```
        y[y<t0] = rlnorm(n, meanlog, sdlog)[y<t0]
    y
}
```

Now, let's try to compute the 0.4 quantile for the above random numbers using the `quantile` function in R. Here we generate fairly large sample sizes of $n = 100,000$ and $n = 1,000,000$ to get a more accurate estimate of the 0.4 quantile.

```
# set seed for reproducibility
set.seed(123)

# generate 100,000 random numbers from truncated log-normal distribution
y = rtrunc_lnorm(n = 100000, meanlog = 1, sdlog = 1.2, t0 = 2)

# calculate the 0.4 quantile
quantile(y, 0.4)
```

```
##      40%
## 4.179444
```

```
# generate 1,000,000 random numbers from truncated log-normal distribution
y = rtrunc_lnorm(n = 1000000, meanlog = 1, sdlog = 1.2, t0 = 2)

# calculate the 0.4 quantile
quantile(y, 0.4)
```

```
##      40%
## 4.177875
```

This is very close to the value of 4.172 obtained from the algorithm in question C3.

D: Cox's partial likelihood with a time-varying effects

Q1. Let the right censored data tuple $(t_i, \delta_i, x_i(.))$ for individual $i \in \{1, 2, ..., n\}$, with follow-up from time 0 to time t_i, event indicator δ_i (with value 1 if the event is observed at time t_i, otherwise censored with value 0), and time-varying effects $x_i(t)$. Let the time-varying hazard ratio be $exp(x_i(t)^T\beta)$ for regression coefficients β. Let the risk set $R(t_i)$ be the set of individuals $\{j : t_j \geq t_i\}$. Assume that there are no tied event times. Write out the partial likelihood $L(\beta)$.

The cox proportional hazards model has the hazard at time t given time-varying covariates $x(t)$ as

$h(t|x) = exp(x(t)^T\beta)h_0(t)$, where $h_0(t)$ is the baseline hazard function.

The partial likelihood function $L(\beta)$ is the product over all individuals who experience an event ($\delta_i = 1$) of the conditional probability that that particular individual experiences the event at time t_i, given that the event occurs at that time within the risk set $R(t_i)$.

For an individual i who experiences an event at time t_i, the conditional probability is given by the ratio of their hazard to the sum of the hazards of all individuals in the risk set at time t_i:

$$\frac{h_i(t_i|x_i)}{\sum_{j \in R(t_i)} h_j(t_i|x_i)}$$

$$\Rightarrow \frac{exp(x_i(t_i)^T\beta)h_0(t_i)}{\sum_{j \in R(t_i)} exp(x_j(t_i)^T\beta)h_0(t_i)}$$

$$\Rightarrow \frac{exp(x_i(t_i)^T\beta)}{\sum_{j \in R(t_i)} exp(x_j(t_i)^T\beta)}.$$

Hence, the partial likelihood function $L(\beta)$ is the product of these conditional probabilities over all n individuals for whom an event is observed:

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{exp(x_i(t_i)^T \beta)}{\sum_{j \in R(t_i)} exp(x_j(t_i)^T \beta)} \right)^{\delta_i}$$

where δ_i is the event indicator for individual i, $x_i(t_i)$ is the time-varying effects vector for individual i at observed time t_i, $R(t_i)$ is the risk set at time t_i, β is the vector of regression coefficients.

Note that the x_j in the denominator are evaluated at time t_i .

Q2. Analytically derive the gradient (or score) $\frac{d \log(L)}{d\beta_k}$.

The log partial likelihood function is given by:

$$\log(L(\beta)) = \sum_{i=1}^{n} \delta_i \left(x_i(t_i)^T \beta - \log \left(\sum_{j \in R(t_i)} exp(x_j(t_i)^T \beta) \right) \right).$$

The gradient of the log partial likelihood function with respect to the k-th element of β is:

$$\frac{d \log(L)}{d\beta_k} = \sum_{i=1}^{n} \delta_i \left(x_{ik}(t_i) - \frac{\sum_{j \in R(t_i)} x_{jk}(t_i) exp(x_j(t_i)^T \beta)}{\sum_{j \in R(t_i)} exp(x_j(t_i)^T \beta)} \right).$$

where $x_{ik}(t_i)$ is the k-th element of the time-varying effects vector $x_i(t_i)$ for individual i and

$x_{jk}(t_i)$ is the k-th element of the time-varying effects vector $x_j(t_i)$ for individual j in the risk set $R(t_i)$ at time t_i.

Q3. Let a binary exposure be defined by z_i for individual i and let $x_i(t) = (z_i, z_i t)^T$. Write out a formula for the hazard ratio as a function of time t for those exposed compared with those not exposed.

Considering an exposure z_i, the hazard function for an individual i is given by:

$h_i(t|x_i) = exp(x_i(t)^T \beta) h_0(t)$.

$\Rightarrow exp(z_i \beta_1 + z_i t \beta_2) h_0(t)$.

The hazard ratio at time t for those exposed ($z_i = 1$) compared with those not exposed ($z_i = 0$) is given by:

$\frac{h_i(t|x_i=(1,t)^T)}{h_i(t|x_i=(0,0)^T)} = \frac{exp(\beta_1 + t\beta_2) h_0(t)}{exp(0) h_0(t)} = exp(\beta_1 + t\beta_2)$.

Q4. The following code is used to investigate whether the hazard ratio between distant and localised cancer varies by time. Write out the regression model and carefully interpret the four parameters.

Given the following code:

```
library(survival)
library(biostat3)
transform(biostat3::colon, stage=relevel(stage,"Localised")) |>
    coxph(Surv(surv_mm,status=="Dead: cancer")~stage+tt(stage), data=_,
          tt=function(x,t,...) (x=="Distant")*t/12) |>
    summary()
```

```
## Call:
## coxph(formula = Surv(surv_mm, status == "Dead: cancer") ~ stage +
##     tt(stage), data = transform(biostat3::colon, stage = relevel(stage,
##     "Localised")), tt = function(x, t, ...) (x == "Distant") *
##     t/12)
##
##   n= 15564, number of events= 8369
```

```
##
##                    coef exp(coef) se(coef)      z Pr(>|z|)
## stageUnknown   0.93904   2.55753  0.03777 24.865  < 2e-16 ***
## stageRegional  0.80311   2.23248  0.04105 19.566  < 2e-16 ***
## stageDistant   2.21903   9.19837  0.03561 62.321  < 2e-16 ***
## tt(stage)     -0.12347   0.88385  0.01551 -7.959 1.73e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                exp(coef) exp(-coef) lower .95 upper .95
## stageUnknown      2.5575     0.3910    2.3751    2.7540
## stageRegional     2.2325     0.4479    2.0599    2.4195
## stageDistant      9.1984     0.1087    8.5783    9.8632
## tt(stage)         0.8839     1.1314    0.8574    0.9111
##
## Concordance= 0.727  (se = 0.003 )
## Likelihood ratio test= 5849  on 4 df,   p=<2e-16
## Wald test            = 5310  on 4 df,   p=<2e-16
## Score (logrank) test = 6821  on 4 df,   p=<2e-16
```

From the above code, the cox regression model being fitted can be written as:

$$\mathbf{h(t|stage) = h_0(t)exp(\beta_1 \times stageUnknown + \beta_2 \times stageRegional + \beta_3 \times stageDistant + \beta_4 \times tt(stage))}.$$

where $h(t|stage)$ is the hazard function at time t given the stage of cancer, $h_0(t)$ is the baseline hazard function, $\beta_1, \beta_2, \beta_3, \beta_4$ are the regression coefficients for the stage of cancer, $tt(stage)$ is the time-varying effect for the Distant stage and $stageUnknown, stageRegional, stageDistant$ are indicator variables for the Unknown, Regional, and Distant stages of cancer respectively.

Here, $tt(stage)$ is defined as $(x == Distant) \times t/12$ and Localised is the reference level for the stage of cancer.

The four parameters in question are:

- β_1: The log hazard ratio between the Unknown stage and the Localised stage of cancer.
- β_2: The log hazard ratio between the Regional stage and the Localised stage of cancer.
- β_3: The log hazard ratio between the Distant stage and the Localised stage of cancer at $t = 0$.
- β_4: The time-dependent change in the log hazard ratio between the Distant stage and the Localised stage of cancer per year.

The p-value of each parameter is a very small value ($\ll 0.05$), indicating that the parameters are all statistically significant. This means that all stages of cancer have a significant impact on the hazard of death from cancer compared to the Localised stage and the hazard ratio between Distant and Localised cancer varies over time.

Interpretation of the parameters:

- Patients with Unknown stage cancer have $exp(\beta_1) \sim \mathbf{2.56}$ times higher risk of death from cancer compared to patients with Localised stage cancer.
- Patients with Regional stage cancer have $exp(\beta_2) \sim \mathbf{2.23}$ times higher risk of death from cancer compared to patients with Localised stage cancer.
- Patients with Distant stage cancer have $exp(\beta_3) \sim \mathbf{9.20}$ times higher risk of death from cancer compared to patients with Localised stage cancer.
- The hazard for Distant stage cancer decreases over time at a rate of $1 - exp(\beta_4) \sim 1 - 0.88 = 0.12$, meaning approximately a **12%** reduction in hazard per year.

The 95% confidence interval for the hazard ratio of time varying effect is (0.86, 0.91), which indicates that the hazard for Distant stage cancer decreases by approximately 9% to 14% per year compared to Localised stage cancer with 95% confidence.

E: Data analysis of a randomised controlled trial for hormonal treatment of breast cancer patients in Germany

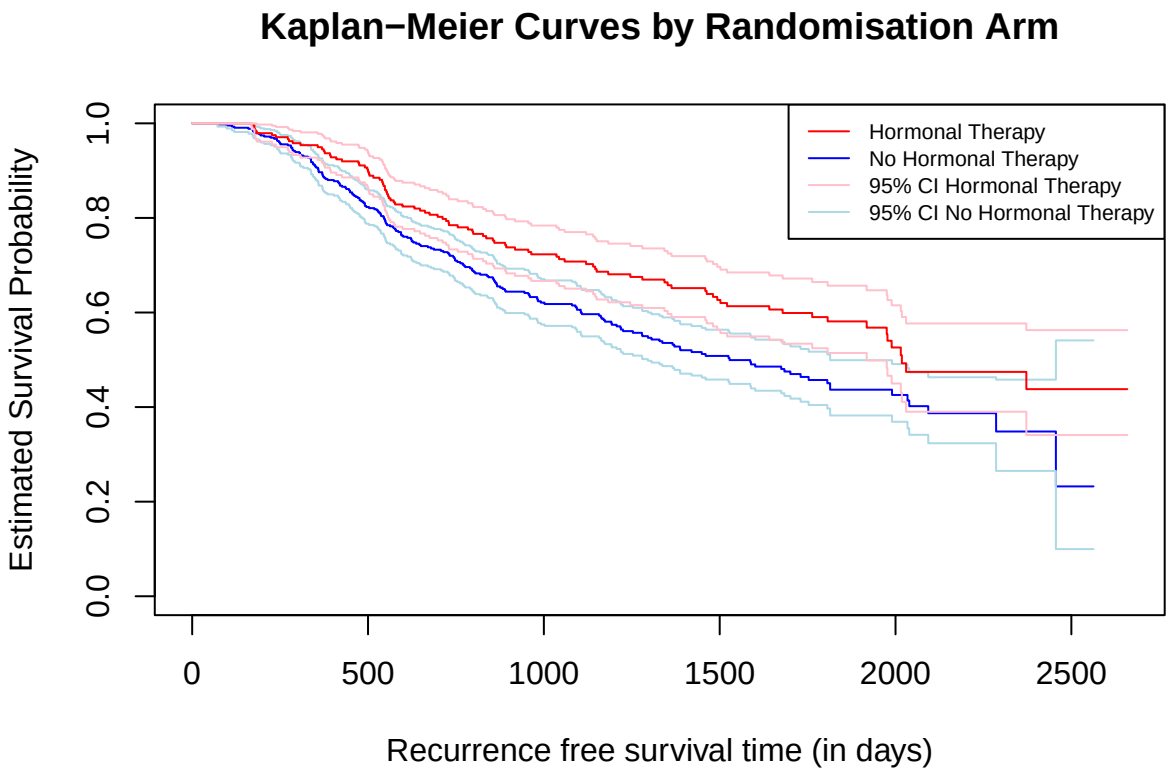
Q1. Plot the Kaplan-Meier curves by randomisation arm, including a legend and appropriate axis labels. Carefully describe the pattern.

To focus on the effect of recurrence in the hormonal therapy with follow-up time rectime and recurrence status indicated by censrec (1 = recurrence, 0 = censored), first we plot the kaplan-meier curves with solely the binary variable hormon (1 = hormonal therapy, 0 = no hormonal therapy) by randomisation arm.

```
# Load the data
library(survival)
library(rstpm2)

data(brcancer)

# Fit the kaplan-meier curves
fit <- survfit(Surv(rectime, censrec) ~ hormon, data = brcancer)

# Preview the summary of the fit for first 5 time points
summary(fit, times = 1:5)
```

```
## Call: survfit(formula = Surv(rectime, censrec) ~ hormon, data = brcancer)
##
##                 hormon=0
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1    440       0        1       0            1            1
##     2    440       0        1       0            1            1
##     3    440       0        1       0            1            1
##     4    440       0        1       0            1            1
##     5    440       0        1       0            1            1
##
##                 hormon=1
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1    246       0        1       0            1            1
##     2    246       0        1       0            1            1
##     3    246       0        1       0            1            1
##     4    246       0        1       0            1            1
##     5    246       0        1       0            1            1
```

```
# Plot the kaplan-meier curves
plot(fit,
     col = c("blue","lightblue", "lightblue", "red", "pink", "pink" ),
     lty = c(1,1),
     conf.int = TRUE, # display confidence intervals
     xlab = "Recurrence free survival time (in days)",
     ylab = "Estimated Survival Probability",
     main = "Kaplan-Meier Curves by Randomisation Arm")

# Add a legend
legend("topright",
       legend = c( "Hormonal Therapy", "No Hormonal Therapy",
                   "95% CI Hormonal Therapy", "95% CI No Hormonal Therapy"),
       col = c("red", "blue", "pink", "lightblue"),
       lty = c(1,1),
       cex = 0.7)
```



We see the following pattern in the Kaplan-Meier curves:

- **Initial Survival Probability:** At time zero, both curves start at survival probability 1.0 as no patients have experienced recurrence yet.
- **Separation of Curves:** The survival probability for patients receiving hormonal therapy remains higher than those not receiving hormonal therapy throughout the follow-up period. This indicates that hormonal therapy is associated with a lower risk of recurrence.
- **Nature of the Curves:** The Kaplan Meier estimate remains constant between events and drops only at observed recurrence times, leading to stepwise changes in the curves. The curves show a decreasing trend in survival probability over time as more patients experience recurrence.
- **Confidence Intervals:** The lighter lines around the Kaplan-Meier curves represent the 95% confidence intervals for the survival probabilities. The confidence intervals are wider at later time points due to fewer patients being at risk. The upper bound of no hormonal therapy arm and the lower bound of hormonal therapy arm jump above and below each other throughout the follow-up period.
- **Drop in Survival Probability:** The survival probability drops sharply at 2500 days for the no hormonal therapy arm to 0.2 survival probability, indicating fewer patients surviving without recurrence after 7 years. Whereas, the hormonal therapy arm shows a more gradual decrease in survival probability at around 500 and 2000 days.

The hormonal therapy arm shows a consistently higher survival probability compared to the no hormonal therapy arm, indicating a beneficial effect of hormonal therapy in reducing the risk of recurrence in breast cancer patients.

Q2. Fit a Cox regression model, adjusting for hormonal treatment. Write out the regression model, defining any notation. Describe your findings, including the estimand of choice to compare those on hormonal treatment compared with those not.

In a Randomised controlled trail, the hormonal treatment is decided randomly irrespective of the patient's characteristics. Although, the number of patients are only 686, the groups are not guaranteed to be perfectly balanced in terms of other covariates. So, in order to assess the potential confounding, let us look at the descriptive statistics of the patients in the two groups.

```r
# Split into two groups of hormonal therapy and no hormonal therapy
brcancer_hormon_0 <- subset(brcancer, hormon == 0)
brcancer_hormon_1 <- subset(brcancer, hormon == 1)

# View the summary of both groups
summary(brcancer_hormon_0)
```

```
##        id            hormon        x1              x2              x3
##  Min.   :  1.0   Min.   :0   Min.   :21.00   Min.   :1.000   Min.   :  3.00
##  1st Qu.:159.8   1st Qu.:0   1st Qu.:45.00   1st Qu.:1.000   1st Qu.: 20.00
##  Median :340.0   Median :0   Median :50.00   Median :1.000   Median : 25.00
##  Mean   :338.0   Mean   :0   Mean   :51.06   Mean   :1.475   Mean   : 29.62
##  3rd Qu.:518.2   3rd Qu.:0   3rd Qu.:59.00   3rd Qu.:2.000   3rd Qu.: 35.00
##  Max.   :686.0   Max.   :0   Max.   :80.00   Max.   :2.000   Max.   :120.00
##        x4              x5               x6             x7
##  Min.   :1.000   Min.   : 1.000   Min.   :   0   Min.   :  0.00
##  1st Qu.:2.000   1st Qu.: 1.000   1st Qu.:   7   1st Qu.:  8.00
##  Median :2.000   Median : 3.000   Median :  32   Median : 32.00
##  Mean   :2.143   Mean   : 4.943   Mean   : 102   Mean   : 79.72
##  3rd Qu.:3.000   3rd Qu.: 7.000   3rd Qu.: 130   3rd Qu.: 92.25
##  Max.   :3.000   Max.   :51.000   Max.   :1600   Max.   :898.00
##     rectime          censrec            x4a              x4b
##  Min.   :   8.0   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 547.8   1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:0.0000
##  Median : 967.0   Median :0.0000   Median :1.0000   Median :0.0000
##  Mean   :1059.7   Mean   :0.4659   Mean   :0.8909   Mean   :0.2523
##  3rd Qu.:1573.0   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :2563.0   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##       x5e
##  Min.   :0.002198
##  1st Qu.:0.431710
##  Median :0.697676
##  Mean   :0.634042
##  3rd Qu.:0.886921
##  Max.   :0.886921
```

```r
summary(brcancer_hormon_1)
```

```
##        id            hormon        x1              x2              x3
##  Min.   :  2.0   Min.   :1   Min.   :32.00   Min.   :1.00   Min.   :  4.0
##  1st Qu.:200.8   1st Qu.:1   1st Qu.:50.00   1st Qu.:2.00   1st Qu.: 20.0
##  Median :355.5   Median :1   Median :58.00   Median :2.00   Median : 25.0
##  Mean   :353.3   Mean   :1   Mean   :56.62   Mean   :1.76   Mean   : 28.8
##  3rd Qu.:503.8   3rd Qu.:1   3rd Qu.:63.00   3rd Qu.:2.00   3rd Qu.: 35.0
##  Max.   :683.0   Max.   :1   Max.   :80.00   Max.   :2.00   Max.   :100.0
##        x4              x5              x6               x7
##  Min.   :1.000   Min.   : 1.00   Min.   :   0.00   Min.   :   0.0
##  1st Qu.:2.000   1st Qu.: 1.00   1st Qu.:   7.25   1st Qu.:   9.0
##  Median :2.000   Median : 3.00   Median :  35.00   Median :  46.0
##  Mean   :2.069   Mean   : 5.13   Mean   : 124.29   Mean   : 125.8
##  3rd Qu.:2.000   3rd Qu.: 7.00   3rd Qu.: 133.00   3rd Qu.: 182.5
##  Max.   :3.000   Max.   :36.00   Max.   :2380.00   Max.   :1144.0
##     rectime          censrec            x4a              x4b
##  Min.   :  15.0   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
```

```
##  1st Qu.: 695.8   1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:0.0000
##  Median :1220.5   Median :0.0000   Median :1.0000   Median :0.0000
##  Mean   :1240.3   Mean   :0.3821   Mean   :0.8659   Mean   :0.2033
##  3rd Qu.:1818.0   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :2659.0   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##       x5e
##  Min.   :0.0133
##  1st Qu.:0.4317
##  Median :0.6977
##  Mean   :0.6196
##  3rd Qu.:0.8869
##  Max.   :0.8869
```

The major differences in the mean and median are observed for $x6$ and $x7$ between the two groups, i.e., the progesterone receptor and the estrogen receptor. So, we can consider adjusting for these covariates in the Cox regression model, along with the hormonal treatment.

$h(t|hormon, x6, x7) = h_0(t)exp(\beta_1 \times hormon + \beta_2 \times x6 + \beta_3 \times x7)$

where $h(t|hormon, x6, x7)$ is the hazard function at time t given hormonal treatment and covariates $x6$ and $x7$, $h_0(t)$ is the baseline hazard function (for the reference group: $hormon = 0$, $x6 = 0$, $x7 = 0$), $hormon$ is the indicator variable for hormonal therapy where $hormon = 1$ if patient has received hormonal therapy and $hormon = 0$ otherwise, $x6$ is the progesterone receptor in fmol, $x7$ is the estrogen receptor in fmol, and $\beta_1, \beta_2, \beta_3$ are the regression coefficients for hormonal treatment, progesterone receptor, and estrogen receptor respectively.

The Cox regression model is fitted as follows:

```r
# Fit the Cox regression model
cox_model <- coxph(Surv(rectime, censrec) ~ hormon + x6 + x7, data = brcancer)

# Display the summary of the model
summary(cox_model)
```

```
## Call:
## coxph(formula = Surv(rectime, censrec) ~ hormon + x6 + x7, data = brcancer)
##
##   n= 686, number of events= 299
##
##              coef  exp(coef)   se(coef)      z Pr(>|z|)
## hormon -3.408e-01  7.112e-01  1.255e-01 -2.716   0.0066 **
## x6     -2.746e-03  9.973e-01  5.964e-04 -4.605 4.13e-06 ***
## x7      2.703e-05  1.000e+00  4.562e-04  0.059   0.9527
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## hormon    0.7112      1.406    0.5561    0.9095
## x6        0.9973      1.003    0.9961    0.9984
## x7        1.0000      1.000    0.9991    1.0009
##
## Concordance= 0.629  (se = 0.016 )
## Likelihood ratio test= 41.74  on 3 df,   p=5e-09
## Wald test            = 30.68  on 3 df,   p=1e-06
## Score (logrank) test = 28.35  on 3 df,   p=3e-06
```

The Cox regression model estimates the hazard ratio (ψ) for hormonal treatment compared to no hormonal

treatment. The estimand of choice is the hazard ratio, which represents the relative risk of recurrence for patients receiving hormonal treatment compared to those not receiving hormonal treatment. This is equivalent to the exponential of log hazard ratio of hormonal treatment, i.e., the regression coefficient β in the Cox regression model.

The Cox regression model assumes that the hazard of breast cancer recurrence is proportional between the two treatment arms i.e., hazard ratio between groups is constant over time.

The findings from the Cox regression model are as follows:

- **Hazard Ratio for Hormonal Treatment:** The hazard ratio for hormonal treatment compared to no hormonal treatment is estimated to be $exp(\beta_1) = 0.711$. This indicates that the hazard of breast cancer recurrence for patients receiving hormonal therapy is 28.9% lower than for those not receiving hormonal therapy, after adjusting for hormonal treatment, progesterone receptor and estrogen receptor.
- **Hazard Ratio for Progesterone and Estrogen Receptor:** The hazard ratio for the progesterone receptor ($x6$) is estimated to be $exp(\beta_2) = 0.997$, indicating that for every unit increase in the progesterone receptor, the hazard of recurrence changes by 0.3%. The hazard ratio for the estrogen receptor ($x7$) is estimated to be $exp(\beta_3) = 1.000$, indicating that a unit increase in the estrogen receptor does not change the hazard of recurrence.
- **Statistical Significance:** The p-value associated with the *hormon* parameter is much less than 0.05, indicating that the effect of hormonal treatment on the risk of recurrence is statistically significant. The p-value of $x6$ is also less than 0.05, indicating that the progesterone receptor is also significantly associated with the risk of recurrence. However, the p-value of $x7$ is almost close to 1, indicating that the estrogen receptor is not significantly associated with the risk of recurrence.
- **Confidence Interval for Hormonal Treatment:** The 95% confidence interval suggests that the true hazard ratio for hormonal treatment compared to no hormonal treatment lies between 0.544 and 0.888 with 95% confidence. It does not include 1, further supporting the significance of the hormonal treatment effect. This is shown in the Kaplan-Meier curves where the hormonal therapy arm shows a consistently higher survival probability compared to the no hormonal therapy arm.
- **Model Fit:** The model fits the data reasonably well, as indicated by the small p-values of Likelihood ratio test, Wald test, and Score test. The concordance index of 0.629, suggests that the model has moderate predictive power.

In conclusion, the Cox regression model shows that hormonal treatment is associated with a significantly lower risk of breast cancer recurrence compared to no hormonal treatment, after adjusting for progesterone receptor and estrogen receptor. The hazard ratio for hormonal treatment is 0.711, indicating a 28.9% reduction in the hazard of recurrence for patients receiving hormonal therapy. The progesterone receptor was significant in predicting the risk of recurrence, although did not influence the hazard of recurrence. The estrogen receptor was significant in neither predicting the risk of recurrence nor influencing the hazard of recurrence.

Q3. Provide a formal test for proportional hazards by treatment arm. Clearly describe which test, motivate why you chose that test, and describe what the test found.

We are interested in testing whether the effect of hormonal treatment on the hazard of recurrence is constant over time or we need to consider time-varying effects of hormonal treatment. Since, the Cox regression model assumes that the hazard ratio is constant over time, we need to test the proportional hazards assumption to ensure the validity of the model.

We can use the Schoenfeld residuals test, since it is a widely used and straightforward method for assessing the proportional hazards assumption in Cox regression models. The Schoenfeld residuals represent the difference between the observed and expected values of the covariate at each event time, and the test assesses whether these residuals are independent of time.

Null Hypothesis H_0: The effect of hormonal treatment on the hazard of recurrence is constant over time (proportional hazards assumption holds).

Alternate Hypothesis H_1: The hormonal treatment has time-varying effects on the hazard of recurrence (proportional hazards assumption is violated).

For p-value < 0.05, we reject the null hypothesis and conclude that the proportional hazards assumption is violated.

We use the `cox.zph` function in R to perform the Schoenfeld residuals test, which calculates the scaled Schoenfeld residuals and tests for independence of these residuals with time.

```
# Test for proportional hazards using Schoenfeld residuals
schoenfeld_test <- cox.zph(cox_model)

# Display the results of the Schoenfeld residuals test
schoenfeld_test
```

```
##        chisq df     p
## hormon 0.313  1 0.576
## x6     6.207  1 0.013
## x7     5.695  1 0.017
## GLOBAL 9.395  3 0.024
```

The Schoenfeld residuals test results are as follows:

- **Hormonal Treatment:** The p-value for the test of proportional hazards for hormonal treatment is 0.576, which is not statistically significant. Hence, we do not reject the null hypothesis that hormonal treatment is constant over time. This suggests that the effect of hormonal treatment on the hazard of recurrence is consistent over time and supports the proportional hazards assumption.

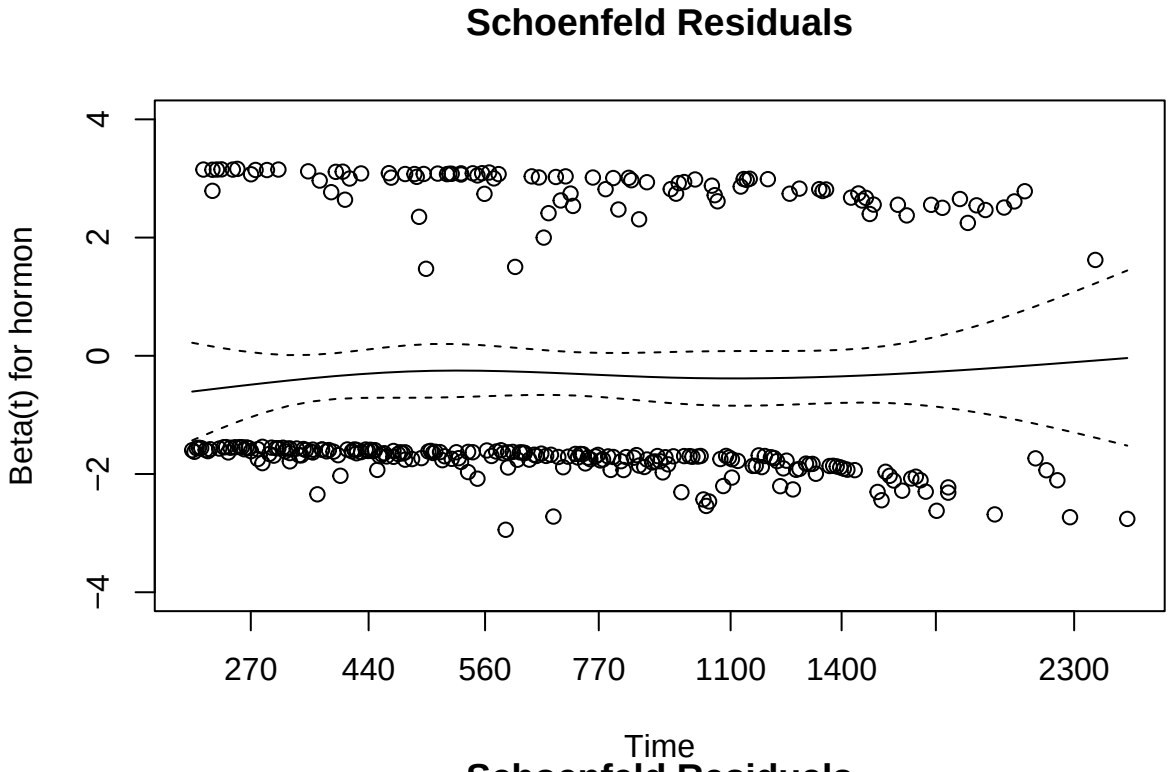
Although, the p-value for $x6$, $x7$ and global are all less than 0.05, indicating that the effect of progesterone receptor and estrogen receptor on the hazard of recurrence, and the model as a whole is not consistent over time. But, we know that they have minimal influence on the hazard of recurrence from the Cox regression model, so we can ignore this violation of proportional hazards assumption.

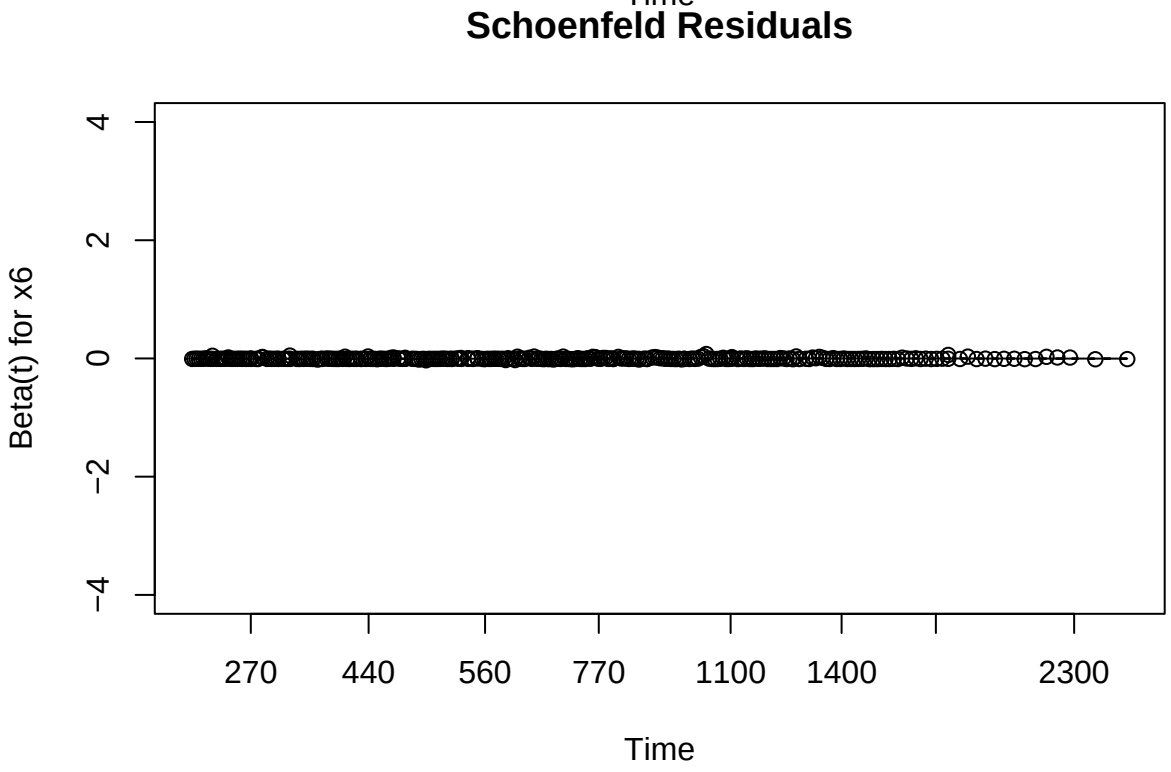
Q4. Provide a plot to graphically evaluation whether there is evidence for proportional hazards. Motivate your choice of method, and describe the method and the results of the evaluation.

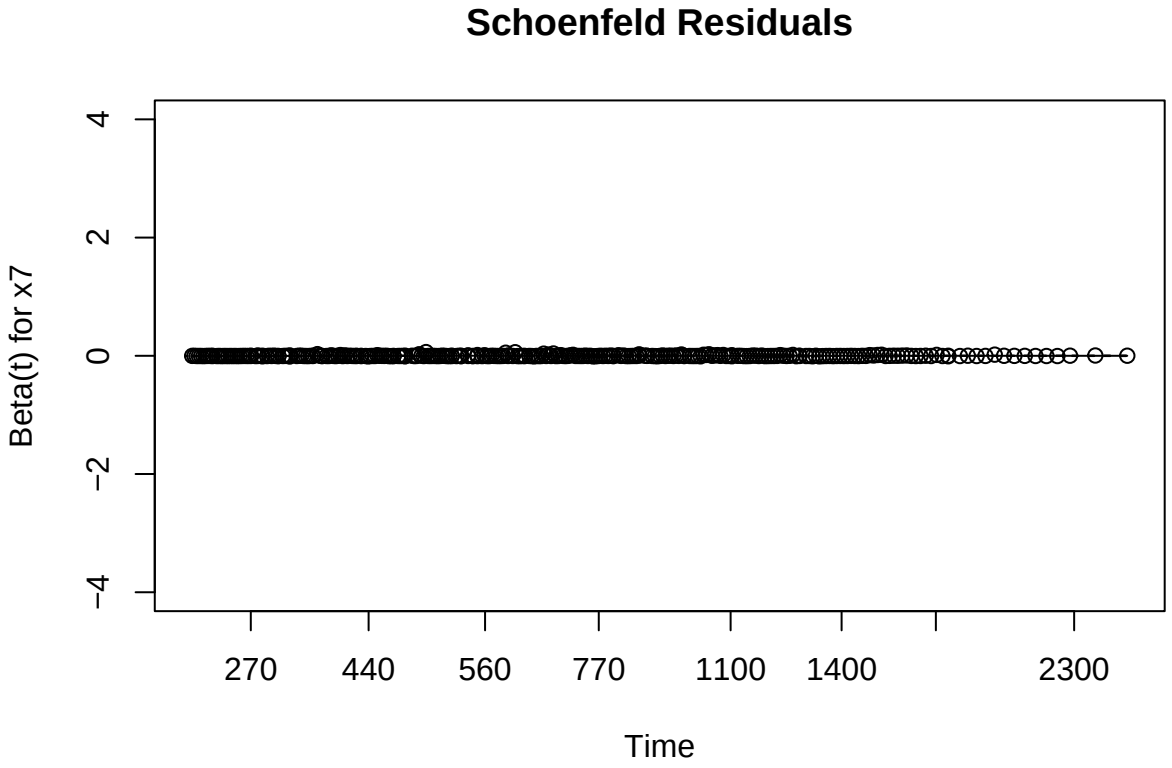
To graphically evaluate whether there is evidence for proportional hazards, we can plot the scaled Schoenfeld residuals against time. This plot can help us visually assess whether the residuals are independent of time, which is a key assumption for proportional hazards.

The plot shows the residuals as a function of time. If the curve shows any systematic pattern or trend, it suggests that the proportional hazards assumption may be violated. On the other hand, if it is horizontal and centered around zero, it indicates that the proportional hazards assumption holds. This test helps us visually understand the pattern of hormonal treatment effect over time.

```
# Plot the Schoenfeld residuals against time
par(mfrow = c(1,1))
for(var in 1:3) {
  plot(schoenfeld_test, var = var, resid = TRUE, se = TRUE,
       main = "Schoenfeld Residuals", ylim = c(-4,4))
}
```


Schoenfeld Residuals
Beta(t) for hormon
4
2
0
−2
−4
270
440
560
770
1100
1400
2300
Time


Schoenfeld Residuals
Beta(t) for x6
4
2
0
−2
−4
270
440
560
770
1100
1400
2300
Time



Hormonal Treatment:

- The curve is very close to horizontal and centered around zero, which supports the proportional hazards assumption.
- The confidence interval bands around the residuals are narrow and do not show any systematic pattern, further supporting the assumption of proportional hazards.

Progesterone Receptor ($x6$) and Estrogen Receptor ($x7$):

- The curves for the progesterone receptor and estrogen receptor coincide with the horizontal line at zero, indicating that their influence on the hazard of recurrence is consistent over time.

In conclusion, the graphical evaluation of the Schoenfeld residuals against time shows the log-hazard ratio for hormonal treatment is constant over time, indicating that the proportional hazards assumption holds for the Cox regression model by treatment arm. This suggests that the effect of hormonal treatment on the hazard of recurrence is consistent over time, as well as for the progesterone and estrogen receptors, and time varying effects are not required in the model. This supports the validity of the Cox regression model for assessing the effect of hormonal treatment on the risk of breast cancer recurrence.

F: Analysis plan for a randomised controlled trial

Q1. Write an outline of how to analyse this study. The outline should include: the estimand of interest; the estimator for that estimand; how you will assess potential confounding; how you will model for potential confounding; whether and how to assess for proportional hazards; and a description of the tables and figures used for reporting. Your analysis plan should provide some motivation for your choices.

For the research question of *Does low-dose aspirin improve survival for patients diagnosed with colorectal cancer with a particular genetic signature?*, we are interested in estimating the causal effect of low-dose aspirin on the survival of patients with a particular genetic signature diagnosed with colorectal cancer over five years

of follow-up. The patients with the genetic signature are randomly assigned to either the experimental arm (low-dose aspirin) or the control arm (no low-dose aspirin).

Analysis Plan:

- **Estimand of Interest:** Causal effect of low-dose aspirin on the survival of patients can be measured by comparing the cause-specific death due to colorectal cancer between the experimental arm (low-dose aspirin) and the control arm (no low-dose aspirin) for patients with a particular genetic signature. This can be quantified using the Hazard Ratio, which is the ratio of the hazard of death due to colorectal cancer in the experimental arm to the hazard in the control arm. The hazard ratio is easy to interpret and provides a measure of the relative risk of death due to colorectal cancer between the two arms. Hazard ratio less than 1 indicates a beneficial effect of low-dose aspirin on survival.
- **Estimator:** The hazard ratio is primarily estimated using the Cox proportional hazards model. This model is mainly used for time-to-event data, and can adjust for potential confounders, handle censoring data, and assess the effect of low-dose aspirin on cause-specific death due to colorectal cancer. The Cox model is suitable for this study as it can model time-varying covariates, adjust for potential confounders, does not require assumption about underlying survival distribution and assess the effect of low-dose aspirin on survival.
- **Assessment of Potential Confounding:** Since this is a RCT as well, we know that the randomization process helps in balancing the covariates between the treatment and control arms. However, we can assess the balance of covariates between the two arms using descriptive statistics. Once, we understand the characteristics of the study population between the two arms, we can decide which covariates to adjust for in the Cox model.
- **Modeling for Potential Confounding:** We will adjust for potential confounders by including them as covariates in the Cox regression model. The regression model will look like:

$h(t|aspirin, X) = h_0(t)exp(\beta \times aspirin + \gamma \times X)$,

where $h(t|aspirin, X)$ is the hazard function at time t given aspirin treatment and covariates X, such as age, sex, cancer stage etc. $h_0(t)$ is the baseline hazard function for no aspirin treatment and the reference levels of all the covariates, β is the regression coefficient for aspirin treatment, $aspirin$ is the indicator variable for low-dose aspirin treatment, where $aspirin = 1$ if patient has received low-dose aspirin and $aspirin = 0$ otherwise, and γ are the regression coefficients for covariates X.

By adjusting for confounders, we want to estimate the causal effect of low-dose aspirin on cause-specific death due to colorectal cancer, conditional on the other covariates.

- **Assessment for Proportional Hazards:** We will assess the proportional hazards assumption using the Schoenfeld residuals test. We can test for proportional hazards by examining the scaled Schoenfeld residuals against time. A non-significant p-value indicates that the proportional hazards assumption holds. If the proportional hazards assumption is violated, we will consider time-varying effects as a part of the model.
- **Tables and Figures for Reporting:**
  - **Table 1:** Descriptive statistics of the study population, including baseline characteristics by treatment arm. This table will include the summary of patient demographics, cancer stage, and other relevant covariates, to help understand the characteristics of the study population.
  - **Table 2:** Results of the Cox regression model, including hazard ratio, 95% confidence interval, and p-value for the effect of low-dose aspirin on cause-specific death due to colorectal cancer. This summary will show us the individual hazard ratios of each of the potential confounders conditional on the other covariates and their significance level in the model.
  - **Figure 1:** Kaplan-Meier curves for the experimental and control arms, showing the survival probability over time. This will help us visualize the difference in survival between the two arms, and assess the effect of low-dose aspirin on survival on a high level.

– **Figure 2:** Schoenfeld residuals plot to assess the proportional hazards assumption. This plot will help us evaluate whether the hazard ratio for low-dose aspirin is constant over time, and whether time varying effects need to be considered in the model.

Hence, with the help of the Cox proportional hazards model, we can estimate the causal effect of low-dose aspirin on cause-specific death due to colorectal cancer, while adjusting for potential confounders and assessing the proportional hazards assumption.

Q2. Finally, discuss whether non-collapsibility is an issue for the chosen estimator.

Yes, non-collapsibility can be an issue for the chosen estimator, the Cox proportional hazards model. Hazard ratio in Cox proportional hazards model is a non-collapsible measure of association. This means that the marginal hazard ratio, i.e., the hazard ratio seen when not conditioned on any other variables, can be different from the conditional hazard ratio, i.e., the hazard ratio seen when conditioned on other variables. This can be seen even when the other variables are not confounders.

We estimate the conditional hazard ratio in the Cox model when adjusted for confounders X as:

$h(t|aspirin, X) = h_0(t)exp(\beta \times aspirin + \gamma \times X)$.

In the context of the study, non-collapsibility can be an issue if we are interested in the marginal effect of low-dose aspirin on cause-specific death due to colorectal cancer, independent of other covariates, such as the patient's age, sex and their stage of cancer, i.e., $h(t|aspirin)$.

We can address the non-collapsibility by using regression standardisation to calculate the marginal effect, which can provide a better interpretative measure of the effect of low-dose aspirin from the conditional model. This method can help us estimate the average causal effect of low-dose aspirin on cause-specific death due to colorectal cancer, independent of other covariates.

We can also use Aalen's additive hazards model, which estimates the additive effect of low-dose aspirin on the cumulative hazard function, providing a collapsible measure of the marginal effect of the treatment. But, in this case, the estimand is not the hazard ratio, but the difference in the cumulative hazard function between the treatment and control arms, which is harder to interpret than the hazard ratio.

In practice, the Cox model is widely used due to its flexibility and ease of interpretation, it is important to be aware of the non-collapsibility issue when interpreting the hazard ratios from the model.

I am declaring that I have used generative artificial intelligence (GAI) to assist me in completing this assignment in the form of Github co-pilot to improve my summarization and writing skills.

This assignment took me approximately 25 hours to complete.