Survival Analysis with Applications in Medicine: Take-home examination

2025-03-21

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.

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

Here, $S(t) = E_X[S(t \mid X)]$, which will also be 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.

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

Similarly, $S(t) = E_X[S(t \mid X)]$, which will also be 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 $S(t|x) = \exp(-\lambda t^k \exp(-k\tilde{\beta}x))$.

Comparing 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.

Hence, the log-likelihood in terms of Survival function $S(t)$ at time t 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.

Yes, we can express the interval-censored data using the `Surv` function from the `survival` package.

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.

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

```r
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 $Q(p) = S^{-1}(1 - p) \rightarrow Q(1 - p) = S^{-1}(p)$.

Hence, for a y, we can say that $S^{-1}(y) = Q(1 - y)$,

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

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

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 $p = 0.4$.

```r
#' @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 for large sample sizes of $n = 10,000$ and $n = 100,000$.

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

# generate 10,000 random numbers from truncated log-normal distribution
y = rtrunc_lnorm(n = 10000, meanlog = 1, sdlog = 1.2, t0 = 2)

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

```
##      40%
## 4.149261
```

```
# 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%
## 4.164862
```

We can see that the 0.4 quantile is approximately 4.149 and 4.165 by randomly sampling t values from a truncated log-normal distribution of T for $n = 10,000$ and $n = 100,000$ respectively. 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:

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

Substituting the above form of the hazard function, the baseline hazard $h_0(t_i)$ cancels out from the numerator and the denominator:

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

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

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

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")) |>
```

```r
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, and $\beta_1, \beta_2, \beta_3, \beta_4$ are the regression coefficients for the stage of cancer.

The model includes three stages of cancer, such as Unknown, Regional, Distant, and a time-varying effect `tt(stage)`, defined as $(x == Distant) \times t/12$, which is only applicable when the stage is Distant.

Here, 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, much less than 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.

The time-varying effect for the Distant stage is also significant, indicating that the hazard ratio between Distant and Localised cancer varies over time.

Let us interpret these parameters using the hazard ratios mentioned in the summary:

- Patients with Unknown stage cancer have $exp(\beta_1) \sim$ **2.56** times higher risk of death from cancer compared to patients with Localised stage cancer.
- Patients with Regional stage cancer have $exp(\beta_2) \sim$ **2.23** times higher risk of death from cancer compared to patients with Localised stage cancer.
- Patients with Distant stage cancer have $exp(\beta_3) \sim$ **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.

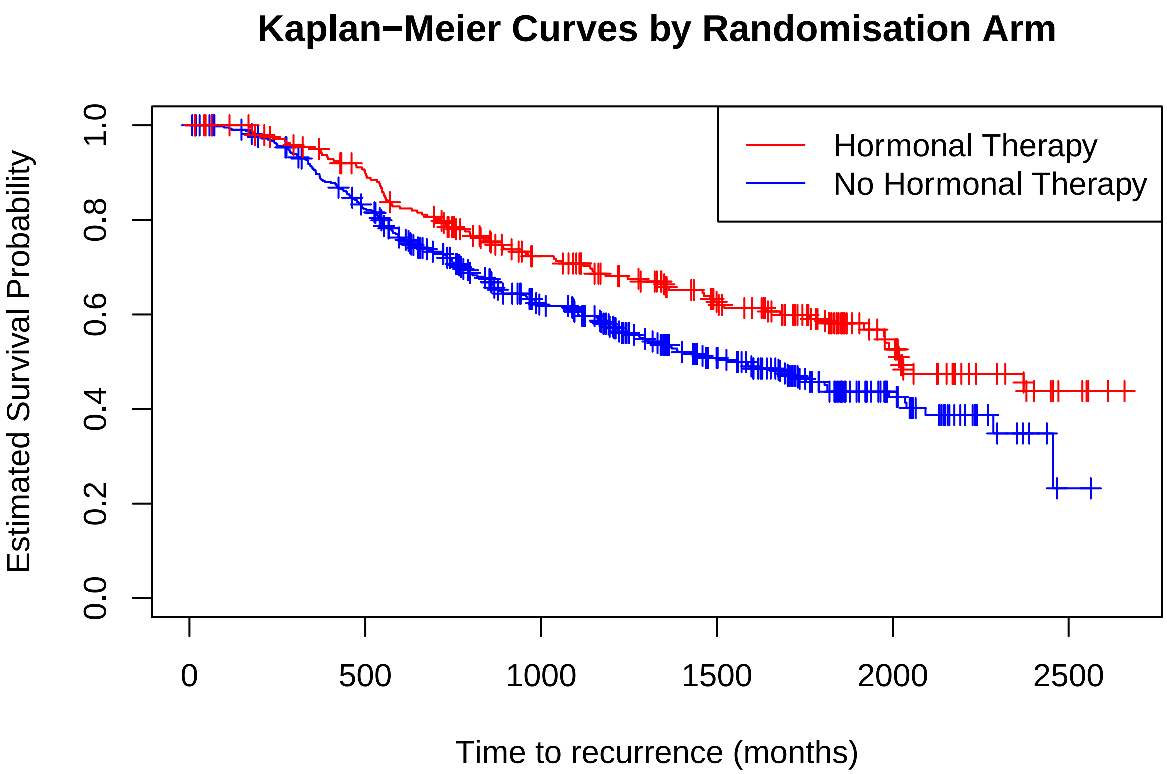
To focus on the effect of recurrence in the hormonal therapy with follow-up time rectime and recurrence status indicated by censrec (1 = recurrence, 0 = censored), first we plot the kaplan-meier curves by randomisation arm.

```
# 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", "red"),
     lty = c(1,1),
     mark.time = TRUE, # display censored events
     xlab = "Time to recurrence (months)",
     ylab = "Estimated Survival Probability",
     main = "Kaplan-Meier Curves by Randomisation Arm")

# Add a legend
legend("topright",
       legend = c( "Hormonal Therapy", "No Hormonal Therapy"),
       col = c("red", "blue"),
       lty = c(1,1))
```



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.
- **Censoring:** Censored events are marked by small vertical ticks, indicating patients who were censored, i.e., did not experience recurrence during the follow-up period.

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.

To fit a Cox regression model adjusting for hormonal treatment, we consider the following regression model:

$h(t|hormon) = h_0(t)exp(\beta \times hormon)$

where $h(t|hormon)$ is the hazard function at time t given hormonal treatment, $h_0(t)$ is the baseline hazard function (for the reference group: $hormon = 0$), $hormon$ is the indicator variable for hormonal therapy where $hormon = 1$ if patient has received hormonal therapy and $hormon = 0$ otherwise and β is the regression coefficient for hormonal treatment.

The Cox regression model is fitted as follows:

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

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

```
## Call:
## coxph(formula = Surv(rectime, censrec) ~ hormon, data = brcancer)
##
##   n= 686, number of events= 299
##
##           coef exp(coef) se(coef)      z Pr(>|z|)
## hormon -0.3640    0.6949   0.1250 -2.911   0.0036 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## hormon    0.6949      1.439    0.5438    0.8879
##
## Concordance= 0.543  (se = 0.014 )
## Likelihood ratio test= 8.82  on 1 df,   p=0.003
## Wald test            = 8.47  on 1 df,   p=0.004
## Score (logrank) test = 8.57  on 1 df,   p=0.003
```

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 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:** The hazard ratio for hormonal treatment compared to no hormonal treatment is estimated to be $\psi = exp(\beta) = 0.695$. This indicates that the hazard of breast cancer recurrence for patients receiving hormonal therapy is 30.5% lower than for those not receiving hormonal therapy, after adjusting for hormonal treatment in the model.
- **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.
- **Confidence Interval:** 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 0.543 indicates that the model has poor predictive accuracy, but this is expected given the simplicity of the model and the limited number of covariates included. This tells us that the Cox regression model cannot predict the survival of patients with high accuracy.

In conclusion, we can say that the Cox regression model fitted to data from 686 patients with 299 recurrence events, using recurrence status and follow up time, suggests that hormonal treatment is associated with a significantly lower risk of breast cancer recurrence compared to no hormonal treatment, with an estimated hazard ratio of 0.695.

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.

To test for the proportional hazards assumption by treatment arm, we can use the Schoenfeld residuals test. The Schoenfeld residuals test is a widely used and straightforward method for assessing the proportional hazards assumption in Cox regression models.

This test examines whether the scaled Schoenfeld residuals for a given covariate (*hormon*) are independent of time. If they are not independent of time, it suggests that the hazard ratio for that covariate is not constant over time, violating the proportional hazards assumption.

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
cox_model <- coxph(Surv(rectime, censrec) ~ hormon, data = brcancer)
schoenfeld_test <- cox.zph(cox_model)

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

```
##        chisq df    p
## hormon 0.227  1 0.63
## GLOBAL 0.227  1 0.63
```

The Schoenfeld residuals test results are as follows:

- **Chi-square test statistic:** The chi-square test statistic is 0.227 with 1 degree of freedom, because of one covariate in the model.
- **Hormonal Treatment:** The p-value for the test of proportional hazards for hormonal treatment is 0.63, which is not statistically significant. Hence, we do not reject the null hypothesis.
- **Global Test:** The global test for the proportional hazards assumption has a p-value of 0.63, which is not statistically significant. This suggests that the proportional hazards assumption holds for the Cox regression model by treatment arm.

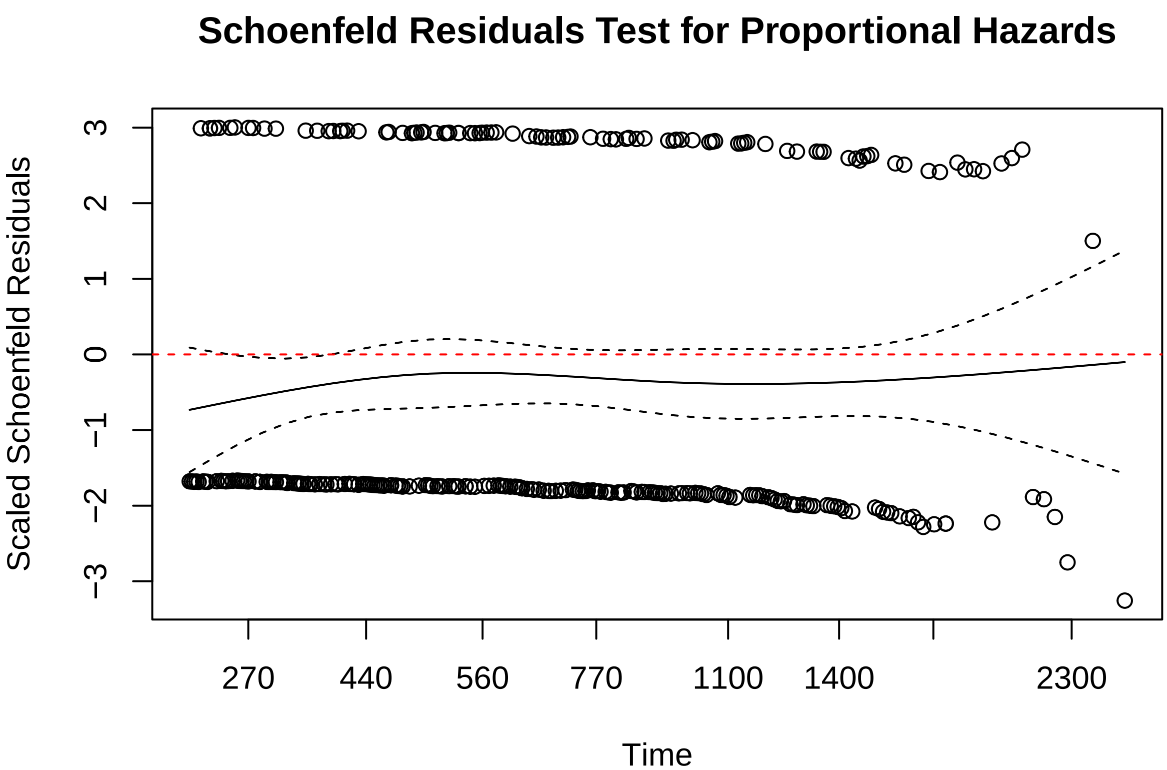
The non-significant p-values for both the global test and the test for hormonal treatment suggest that the proportional hazards assumption holds for the Cox regression model by treatment arm. This means that the hazard ratio for hormonal treatment is constant over time, supporting the validity of the Cox regression model.

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, with a horizontal line at zero indicating that the residuals are independent of time. If the line is not horizontal, it suggests that the proportional hazards assumption is violated, and the hazard ratio is not constant over time, else if the line is horizontal, it indicates that the proportional hazards assumption holds.

```
# Plot the Schoenfeld residuals against time
plot(schoenfeld_test,
     var = 1,
     resid = TRUE, # plot residuals
     se = TRUE,    # plot confidence bands
     xlab = "Time",
     ylab = "Scaled Schoenfeld Residuals",
     main = "Schoenfeld Residuals Test for Proportional Hazards")
abline(h = 0, col = "red", lty = 2) # add a horizontal line at zero
```



- The solid line is 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.

In conclusion, the graphical evaluation of the Schoenfeld residuals against time shows no evidence of violation of the proportional hazards assumption for the Cox regression model by treatment arm. The residuals are independent of time, indicating that the hazard ratio for hormonal treatment is constant over time.

F: Analysis plan for a randomised controlled trial

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 auto-complete my explanations.

This assignment took me approximately 13.5 hours to complete.