# Regresi *Cox Proportional Hazard* dan *Partial Likelihood* pada Data *Veteran*


Disusun oleh :

Ilham Aziz Ramadhan (2206826684)

Maryesta Apriliani S. (2206051531)

Rachmania Azzahra Salsabila (2206825870)

Rahmat Affandi (2206051645)

Vanny Khairunnisaa (2206051506)

**Fakultas Matematika dan Ilmu Pengetahuan Alam**

**Universitas Indonesia**


**2024**

# DAFTAR ISI

# BAB I PENDAHULUAN

## A. Latar Belakang

Kanker paru adalah tumor ganas paru yang berasal dari saluran napas atau epitel bronkus yang ditandai dengan pertumbuhan sel yang tidak normal, tidak terbatas, dan merusak sel-sel jaringan normal. Kanker paru merupakan penyebab utama keganasan di dunia dan mencapai hingga 13% dari semua diagnosis kanker (Kemenkes RI, 2016:1). Data World Health Organization (WHO) menyebutkan bahwa sebesar 8,8 juta kematian di tahun 2015 disebabkan oleh kanker. Dari jumlah tersebut, kanker paru tergolong menduduki peringkat tertinggi yaitu sebesar 1,69 juta kematian. Analisis data ketahanan hidup adalah suatu metode yang berhubungan dengan waktu suatu individu / subjek mulai dari awal pengamatan sampai terjadinya kejadian. Salah satu contoh analisis survival dalam bidang kesehatan dan kedokteran yaitu mengetahui faktor dan kombinasi faktor yang menjadi penyebab kematian pada pasien kanker paru-paru. Kami juga dapat memodelkan tingkat kematian akibat kanker paru-paru dengan mempertimbangkan beberapa faktor kondisi kesehatan seseorang.

Terdapat beberapa metode yang digunakan untuk menganalisis data survival. Analisis regresi dapat digunakan untuk menyelesaikan persoalan analisis survival. Analisis regresi merupakan suatu analisis statistika yang memanfaatkan hubungan antara dua atau lebih peubah kuantitatif sehingga salah satu peubah dapat diramalkan dari peubah lainnya. Salah satu analisis regresi yang terkenal untuk menganalisa data survival adalah regresi Cox. Regresi Cox termasuk ke dalam metode semiparametrik yang mana fungsi baseline hazard mengikuti model nonparametrik sedangkan variabel-variabel independennya mengikuti model parametrik. Tujuan dari metode regresi Cox adalah untuk mengetahui hubungan antara waktu survival dengan variabel variabel yang diduga mempengaruhi waktu survival. Regresi Cox dikenal juga dengan istilah regresi Cox proportional hazard karena asumsi proporsional hazard. Asumsi proporsional hazard merupakan asumsi terpenting yang dipenuhi dalam regresi Cox yang berarti seiring berjalannya waktu tingkat kematian (*event*) adalah konstan. Sebelum melakukan analisis regresi Cox - PH, kami melakukan uji strata dimana hal ini perlu dilakukan memastikan bahwa hasil studi tidak dipengaruhi oleh faktor-faktor luar yang tidak diinginkan. Uji

strata merupakan pengembangan dari uji K sampel dimana terdapat tambahan informasi yang lebih detail mengenai faktor pada masing-masing sampel. Hal ini biasanya karena kondisi di dalam sampel masih cukup heterogen. Dengan demikian, ada faktor yang berperan untuk segmentasi di dalam sampel sedemikian sehingga menjadi lebih homogen. Faktor inilah yang menjadi strata. Uji strata dilakukan dengan membagi data ke dalam kelompok-kelompok yang seragam berdasarkan faktor tertentu yang bisa mempengaruhi hasil (disebut variabel perancu). Dengan demikian, kita dapat memastikan bahwa perbandingan antara kelompok-kelompok dalam sebuah studi adil dan tidak dipengaruhi oleh faktor lain yang bisa membingungkan hasilnya.

Kami akan menggunakan subset data untuk membuat fungsi *partial likelihood* secara keseluruhan dimana fungsi *partial likelihood* merupakan fungsi peluang bersama dari data survival tidak tersensor yang biasa dinotasikan dengan $L(\boldsymbol{\beta})$, di mana parameter-parameter $\boldsymbol{\beta}$ tidak diketahui nilainya (Kleinbaum dan Klein, 2005).

Pada kasus ini, kami akan menggunakan metode Cox proportional hazard, melakukan uji strata, dan membuat fungsi *partial likelihood* untuk mengetahui kombinasi faktor-faktor yang menjadi penyebab event kematian pasien akibat kanker paru-paru, serta untuk memodelkan tingkat kematian akibat kanker paru-paru dengan mempertimbangkan kondisi kesehatan seseorang.

## B. Rumusan Masalah

1) Apa saja kombinasi faktor-faktor yang berperan dalam menyebabkan event kematian pada pasien yang menderita kanker paru-paru?
2) Bagaimana memodelkan tingkat kematian pada pasien yang menderita kanker paru-paru?

## C. Tujuan

1) Mengetahui kombinasi faktor-faktor yang menjadi penyebab event kematian pasien akibat kanker paru-paru.
2) Memodelkan tingkat kematian akibat kanker paru-paru dengan mempertimbangkan kondisi kesehatan seseorang.

## D. Informasi Data

Data yang akan kami gunakan adalah dataset ‘veteran’ yang berisi informasi tentang kumpulan data uji coba acak dua pengobatan untuk pasien penyakit kanker paru-paru yang berkaitan dengan studi kanker paru-paru administrasi veteran. Data ini tersedia di RStudio dalam packages ‘survival’, berikut detail mengenai data yang akan kami gunakan.

```
# Load Packages ----
library(survival)
library(ranger)

library(ggplot2)

library(dplyr)

library(ggfortify)

library(survminer)
```

```
# Load Data ----
data(cancer, package="survival")
head(veteran)
```

```
##   trt celltype time status karno diagtime age prior
## 1   1 squamous   72      1    60        7  69     0
## 2   1 squamous  411      1    70        5  64    10
## 3   1 squamous  228      1    60        3  38     0
## 4   1 squamous  126      1    60        9  63    10
## 5   1 squamous  118      1    70       11  65    10
## 6   1 squamous   10      1    20        5  49     0
```

```
# Deskripsi Data ----
str(veteran) # Observasi: 137. Variabel: 8. Ada 3 variabel numerik prediktor:
age, diagtime, karno
```

```
## 'data.frame':    137 obs. of  8 variables:
##  $ trt     : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ celltype: Factor w/ 4 levels "squamous","smallcell",..: 1 1 1 1 1 1 1 1
1 1 ...
##  $ time    : num  72 411 228 126 118 10 82 110 314 100 ...
##  $ status  : num  1 1 1 1 1 1 1 1 1 0 ...
##  $ karno   : num  60 70 60 60 70 20 40 80 50 70 ...
##  $ diagtime: num  7 5 3 9 11 5 10 29 18 6 ...
##  $ age     : num  69 64 38 63 65 49 69 68 43 70 ...
##  $ prior   : num  0 10 0 10 10 0 10 0 0 0 ...
```

```
# Statistika Deskriptif

summary(veteran) # Mean status: 0.9343 (Persentase data Lengkap: 93,43%)
```

```
##       trt            celltype       time           status
##  Min.   :1.000   squamous :35   Min.   :  1.0   Min.   :0.0000
##  1st Qu.:1.000   smallcell:48   1st Qu.: 25.0   1st Qu.:1.0000
##  Median :1.000   adeno    :27   Median : 80.0   Median :1.0000
##  Mean   :1.496   large    :27   Mean   :121.6   Mean   :0.9343
##  3rd Qu.:2.000                  3rd Qu.:144.0   3rd Qu.:1.0000
##  Max.   :2.000                  Max.   :999.0   Max.   :1.0000
##      karno          diagtime          age            prior
##  Min.   :10.00   Min.   : 1.000   Min.   :34.00   Min.   : 0.00
##  1st Qu.:40.00   1st Qu.: 3.000   1st Qu.:51.00   1st Qu.: 0.00
##  Median :60.00   Median : 5.000   Median :62.00   Median : 0.00
##  Mean   :58.57   Mean   : 8.774   Mean   :58.31   Mean   : 2.92
##  3rd Qu.:75.00   3rd Qu.:11.000   3rd Qu.:66.00   3rd Qu.:10.00
##  Max.   :99.00   Max.   :87.000   Max.   :81.00   Max.   :10.00
```

Ukuran dataset: 137 baris

- Skala/Tipe dataset

| No | Variabel | Definisi | Tipe Data |
|---|---|---|---|
| 1 | Trt : *treatment* | Trt (Treatment): informasi yang mencatat jenis intervensi medis yang diterima oleh pasien<br><br>Keterangan pada data :<br>1=standard=regimen diberikan perlakuan sesuai standar lung cancer;<br>2=test=regimen diberikan perlakuan alternatif atau pengobatan baru | Skala Nominal - object |
| 2 | Cell type : Jenis sel | Cell type : jenis atau kategori spesifik dari sel yang membentuk kanker<br><br>Keterangan pada data :<br>1 : squamou atau skuamosa<br>2 : smallcell<br>3 : adeno atau adenokarsimona<br>4 : large | Skala Nominal - object |
| 3 | Time : Waktu bertahan hidup dalam hitungan hari (*survival time*) | Time :<br>Waktu bertahan hidup pasien dalam hitungan hari (survival time) | Skala Nominal - object |
| 4 | Status : Status penyensoran | Status : Status penyensoran bertujuan untuk menganalisis waktu hingga kanker terjadi,<br>Keterangan pada data :<br>0 : tersensor<br>1 : tidak tersensor | Skala Nominal - object |
| 5 | Karno : Skor kinerja | Karno: Karnofsky performance (kemandirian atau kemampuan pasien menjalani aktivitas)<br><br>Keterangan skor karnofsky pada pasien : | Skala Ordinal - object |

| | | | |
|---|---|---|---|
| | Karnofsky (100 = baik) | ● 100: Normal, tidak ada keluhan, tidak ada tanda-tanda penyakit.<br>● 90: Mampu melakukan aktivitas normal, hanya ada sedikit gejala atau tanda-tanda penyakit.<br>● 80: Mampu melakukan aktivitas normal dengan sedikit usaha lebih, beberapa gejala penyakit.<br>● 70: Mampu merawat diri sendiri, tidak dapat melakukan aktivitas normal atau pekerjaan.<br>● 60: Memerlukan bantuan sesekali tetapi mampu merawat sebagian besar kebutuhannya sendiri.<br>● 50: Memerlukan bantuan yang signifikan dan sering serta perawatan medis.<br>● 40: Cacat, memerlukan bantuan khusus dan perawatan medis.<br>● 30: Sangat cacat, rawat inap diperlukan tetapi kematian tidak mendekat.<br>● 20: Sangat sakit, memerlukan rawat inap dan penunjang medis aktif.<br>● 10: Hampir sekarat, proses kematian mendekat, memerlukan perawatan terus-menerus.<br>● 0: Meninggal. | |
| 6 | Diagtime : Bulan sejak diagnosis hingga pengacakan | Diagtime : waktu saat diagnosis kanker pertama kali oleh dokter hingga pengacakan (dalam bulan) | Skala Nominal - object |
| 7 | Usia : Dalam tahun | Usia : usia pasien yang diobservasi (dalam tahun) | Skala Nominal - object |
| 8 | Prior : Terapi sebelumnya | Prior : informasi pasien apakah pernah melakukan terapi sebelumnya atau tidak<br><br>Keterangan pada data : | Skala Nominal - object |

| | | | |
|---|---|---|---|
| | | ● 0 = tidak<br>● 1 = ya | |

Variabel-variabel dalam dataset ‘veteran’ adalah:

- Trt : *treatment*
  - 1 = standar
  - 2 = uji
- Cell type : Jenis sel
  - 1 = skuamosa
  - 2 = sel kecil
  - 3 = adeno
  - 4 = sel besar
- Time : Waktu bertahan hidup dalam hitungan hari (*survival time*)
- Status : Status penyensoran
- Karno : Skor kinerja Karnofsky (100 = baik)
- Diagtime : Bulan sejak diagnosis hingga pengacakan
- Usia : Dalam tahun
- Prior : Terapi sebelumnya
  - 0 = tidak
  - 1 = ya

# BAB II PREPROCESSING DAN EXPLORATORY DATA ANALYSIS

## A. Pre-Processing

Berikut adalah *packages* dan cuplikan dari baris awal data yang digunakan

```
# Load Packages ----
library(survival)
library(ranger)
library(ggplot2)
library(dplyr)
library(ggfortify)
library(survminer)
library(cluster)
library(forestmodel)
library(psych)
library(MASS)
library(ADGofTest)
library(rms)
library(broom)
```

```
> data(cancer, package="survival")
> head(veteran)
  trt celltype time status karno diagtime age prior
1   1 squamous   72      1    60        7  69     0
2   1 squamous  411      1    70        5  64    10
3   1 squamous  228      1    60        3  38     0
4   1 squamous  126      1    60        9  63    10
5   1 squamous  118      1    70       11  65    10
6   1 squamous   10      1    20        5  49     0
```

Dengan *summary* data didapatkan beberapa informasi seperti nilai minimum, median,mean, nilai maksimum, dan nilai kuartil dari tiap-tiap variabel data sebagai berikut.

```
      trt            celltype       time           status          karno      
 Min.   :1.000   squamous :35   Min.   :  1.0   Min.   :0.0000   Min.   :10.00  
 1st Qu.:1.000   smallcell:48   1st Qu.: 25.0   1st Qu.:1.0000   1st Qu.:40.00  
 Median :1.000   adeno    :27   Median : 80.0   Median :1.0000   Median :60.00  
 Mean   :1.496   large    :27   Mean   :121.6   Mean   :0.9343   Mean   :58.57  
 3rd Qu.:2.000                  3rd Qu.:144.0   3rd Qu.:1.0000   3rd Qu.:75.00  
 Max.   :2.000                  Max.   :999.0   Max.   :1.0000   Max.   :99.00  
    diagtime          age             prior       
 Min.   : 1.000   Min.   :34.00   Min.   : 0.00  
 1st Qu.: 3.000   1st Qu.:51.00   1st Qu.: 0.00  
 Median : 5.000   Median :62.00   Median : 0.00  
 Mean   : 8.774   Mean   :58.31   Mean   : 2.92  
 3rd Qu.:11.000   3rd Qu.:66.00   3rd Qu.:10.00  
 Max.   :87.000   Max.   :81.00   Max.   :10.00  
```

Akan dilakukan pengecekan *missing value* pada data didapatkan tidak ada *missing value* pada data yang digunakan.

```
## Cek Missing Value ----
sum(is.na(veteran)) # Tidak ada missing value

## [1] 0
```

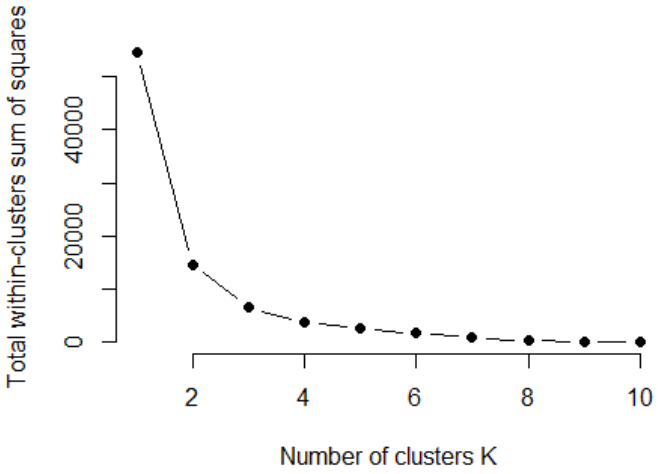
Untuk menghindari kesalahan saat memproses data pada program R, maka nilai setiap variabel akan diubah tipenya menjadi faktor dengan *codes* berikut ini.

```
## Ubah tipe variabel kategorik menjadi factor ----
str(veteran) # trt, prior belum menjadi factor

## 'data.frame':    137 obs. of  8 variables:
##  $ trt     : num  1 1 1 1 1 1 1 1 1 ...
##  $ celltype: Factor w/ 4 levels "squamous","smallcell",..: 1 1 1 1 1 1 1 1
1 1 ...
##  $ time    : num  72 411 228 126 118 10 82 110 314 100 ...
##  $ status  : num  1 1 1 1 1 1 1 1 0 ...
##  $ karno   : num  60 70 60 60 70 20 40 80 50 70 ...
##  $ diagtime: num  7 5 3 9 11 5 10 29 18 6 ...
##  $ age     : num  69 64 38 63 65 49 69 68 43 70 ...
##  $ prior   : num  0 10 0 10 10 0 10 0 0 0 ...

veteran$trt <- as.factor(veteran$trt)
veteran$prior <- as.factor(veteran$prior)
str(veteran) # Lihat hasil

## 'data.frame':    137 obs. of  8 variables:
##  $ trt     : Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...
##  $ celltype: Factor w/ 4 levels "squamous","smallcell",..: 1 1 1 1 1 1 1 1
1 1 ...
##  $ time    : num  72 411 228 126 118 10 82 110 314 100 ...
##  $ status  : num  1 1 1 1 1 1 1 1 0 ...
##  $ karno   : num  60 70 60 60 70 20 40 80 50 70 ...
##  $ diagtime: num  7 5 3 9 11 5 10 29 18 6 ...
##  $ age     : num  69 64 38 63 65 49 69 68 43 70 ...
##  $ prior   : Factor w/ 2 levels "0","10": 1 2 1 2 2 1 2 1 1 1 ...
```

Berdasarkan sumber buku bacaan dikatakan bahwa variabel yang bersifat numerik tidaklah efisien untuk melakukan pemeriksaan fungsi *survival* kecuali jika dilakukan kategorisasi dari variabel numerik tersebut. Kategorisasi variabel numerik dilakukan dengan melakukan *clustering* sehingga variabel numerik dapat berubah menjadi variabel kategorik berdasarkan *cluster*. Untuk mengetahui besar *cluster* tiap variabel yang akan dibentuk, akan digunakan metode Elbow yang diharapkan dapat menghasilkan nilai *k-cluster* optimal yang membagi variabel numerik menjadi variabel kategorik. Diketahui variabel numerik yang ada pada data ini adalah variabel "karno", "diagtime", dan "age".

1) **Mengubah variabel "karno" menjadi "karno_cluster"**

Berikut adalah plot elbow method untuk variabel "karno"

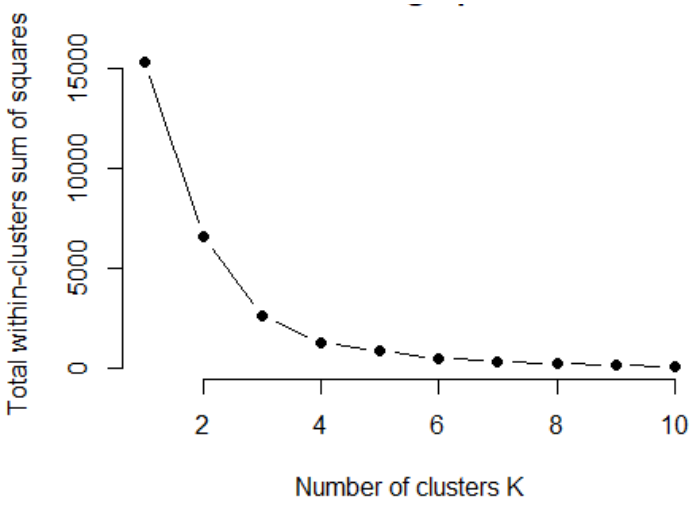



Jika ditinjau dari plot di atas terindikasi bahwa terjadi penurunan pada *cluster* di titik $k = 2$, sehingga dibuatlah variabel kategorik baru "karno_cluster" yang berasal dari variabel "karno" yang dibagi menjadi 2 *cluster*. Variabel "karno_cluster" terbentuk dari 85 data pada *cluster* 1 dan 52 data pada *cluster* 2. Berikut adalah hasil R dari proses *clustering*

```
# Based on the plot, choose the number of clusters
k <- 2
# Perform k-means clustering
set.seed(123) # Setting seed for reproducibility
kmeans_result <- kmeans(karno_data, centers = k, nstart = 25)
# Examine the clustering result
print(kmeans_result)

## K-means clustering with 2 clusters of sizes 85, 52
##
## Cluster means:
##       [,1]
## 1 71.92941
## 2 36.73077
##
## Within cluster sum of squares by cluster:
## [1] 8699.576 5944.231
##  (between_SS / total_SS =  73.2 %)
```

2) **Mengubah variabel "diagtime" menjadi "diagtime_cluster"**

Berikut adalah plot elbow method untuk variabel "diagtime"

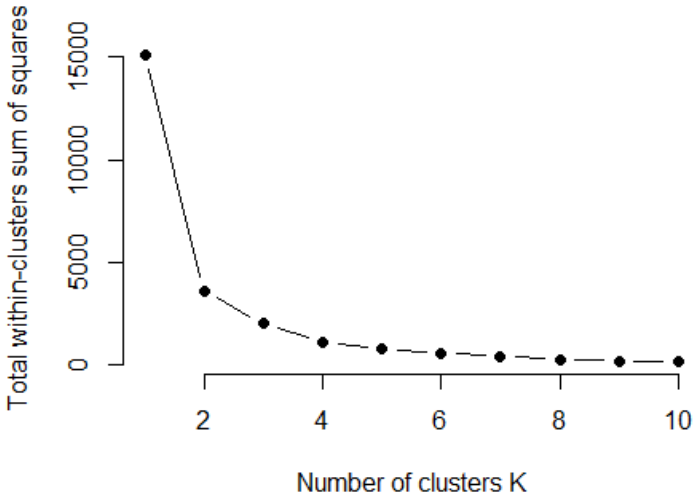



Jika ditinjau dari plot di atas terindikasi bahwa terjadi penurunan pada *cluster* di titik $k = 2$, sehingga dibuatlah variabel kategorik baru "diagtime_cluster" yang berasal dari variabel "diagtime" yang dibagi menjadi 2 *cluster*. Variabel "diagtime_cluster" terbentuk dari 12 data pada *cluster* 1 dan 125 data pada *cluster* 2. Berikut adalah hasil R dari proses *clustering*

```
# Based on the plot, choose the number of clusters
k <- 2
# Perform k-means clustering
set.seed(123) # Setting seed for reproducibility
kmeans_result <- kmeans(diagtime_data, centers = k, nstart = 25)
# Examine the clustering result
print(kmeans_result)

## K-means clustering with 2 clusters of sizes 12, 125
##
## Cluster means:
##       [,1]
## 1 34.58333
## 2  6.29600
##
## Within cluster sum of squares by cluster:
## [1] 4144.917 2410.048
```

**3) Mengubah variabel "age" menjadi "age_cluster"**

Berikut adalah plot elbow method untuk variabel "age"




Jika ditinjau dari plot di atas terindikasi bahwa terjadi penurunan pada *cluster* di titik $k = 2$, sehingga dibuatlah variabel kategorik baru "age_cluster" yang berasal dari variabel "age" yang dibagi menjadi 2 *cluster*. Variabel "age_cluster" terbentuk dari 47 data pada *cluster* 1 dan 90 data pada *cluster* 2. Berikut adalah hasil R dari proses *clustering*

```
# Based on the plot, choose the number of clusters
k <- 2
# Perform k-means clustering
set.seed(123) # Setting seed for reproducibility
kmeans_result <- kmeans(age_data, centers = k, nstart = 25)
# Examine the clustering result
print(kmeans_result)
```

```
## K-means clustering with 2 clusters of sizes 47, 90
##
## Cluster means:
##       [,1]
## 1 45.61702
## 2 64.93333
##
## Within cluster sum of squares by cluster:
## [1] 1857.106 1735.600
##  (between_SS / total_SS =  76.2 %)
```

Berikut adalah *output str* pada data setelah seluruh variabel pada data merupakan variabel kategorik dengan tipe data *factor*

```
str(veteran)
```

```
## 'data.frame':    137 obs. of  11 variables:
##  $ trt            : Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1
...
##  $ celltype       : Factor w/ 4 levels "squamous","smallcell",..: 1
1 1 1 1 1 1 1 1 ...
##  $ time           : num  72 411 228 126 118 10 82 110 314 100 ...
```

```
##  $ status          : num  1 1 1 1 1 1 1 1 1 0 ...
##  $ karno           : num  60 70 60 60 70 20 40 80 50 70 ...
##  $ diagtime        : num  7 5 3 9 11 5 10 29 18 6 ...
##  $ age             : num  69 64 38 63 65 49 69 68 43 70 ...
##  $ prior           : Factor w/ 2 levels "0","10": 1 2 1 2 2 1 2 1 1 1
...
##  $ age_cluster     : Factor w/ 2 levels "1","2": 2 2 1 2 2 1 2 2 1 2
...
##  $ diagtime_cluster: Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 1 2 2
...
##  $ karno_cluster   : Factor w/ 2 levels "1","2": 1 1 1 1 1 2 2 1 2 1
...
```

## B. Exploratory Data and Analysis

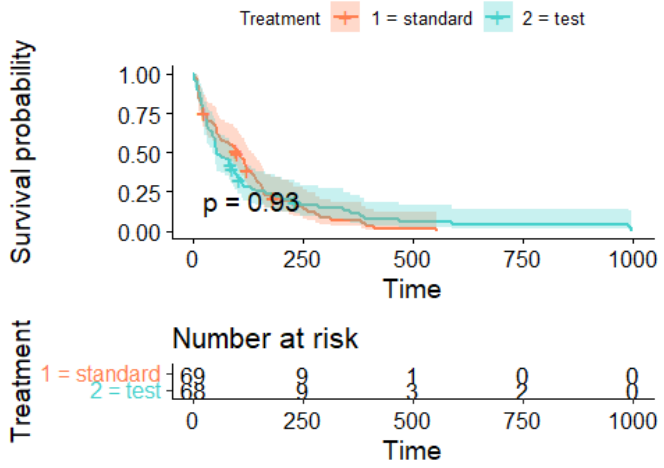
### 1) Statistika Deskriptif

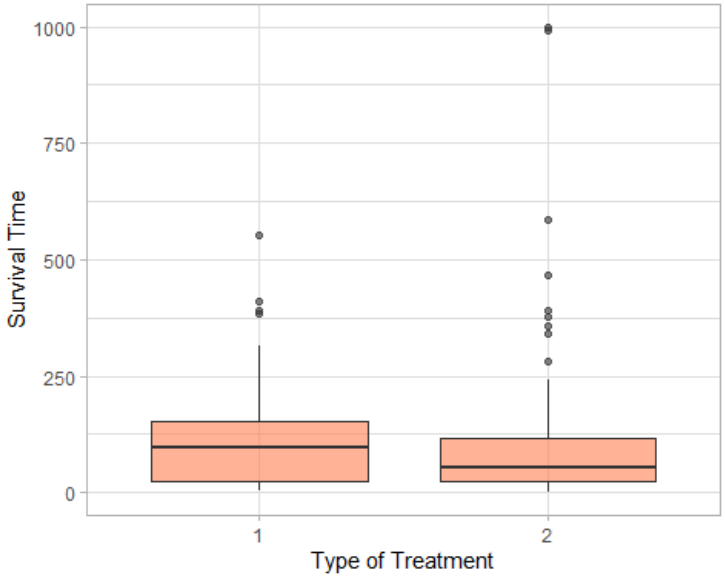
Berdasarkan data veteran yang digunakan, data ini memiliki 11 variabel dan 137 observasi. Berikut adalah informasi singkat mengenai variabel yang digunakan

| Variabel | Distinct | Keterangan |
|---|---|---|
| trt (*treatment*) | 2 | Berisikan $trt_1$ (*treatment standard*) dan $trt_2$ (*treatment alternative*) |
| *celltype* | 4 | Berisikan tipe sel *squamous, smallcell, adeno,* dan *large* |
| *time* | 101 | Berisikan nilai waktu yang berbeda-beda |
| *status* | 2 | Berisikan nilai 0 (tersensor) dan 1 (tidak tersensor) |
| karno | 12 | Berisikan nilai karno yang berbeda-beda dalam rentang (0-100) dengan kelipatan 10 |
| *diagtime* | 28 | Berisikan nilai  yang berbeda-beda |
| *age* | 40 | Berisikan nilai umur yang berbeda-beda |
| *prior* | 2 | Berisikan nilai 0 (tidak pernah melakukan kemo) dan 1 (pernah melakukan kemo) |
| *age_cluster* | 2 | Berisikan $age_1$ dan $age_2$ hasil *clustering* |
| *diagtime_cluster* | 2 | Berisikan $diag_1$ dan $diag_2$ hasil *clustering* |
| *karno_cluster* | 2 | Berisikan $karno_1$ dan $karno_2$ hasil *clustering* |

**2) *Survival Experience* Berdasarkan *Treatment***

Secara berturut-turut berikut adalah plot Estimasi Kaplan Meier, Box-plot, dan Tabel Distribusi dari variable "trt"




```
## Tabel Distribusi
table(veteran$status, veteran$trt)

##
##      1  2
##   0  5  4
##   1 64 64
```

Jika melihat pada *box-plot* dan grafik di atas, terlihat tidak ada perbedaan yang mencolok antara waktu *survival* antar *treatment*. Pada

*box-plot* terlihat beberapa *outlier* yang tidak dibuang untuk analisa lebih lanjut model. Untuk analisis lebih lanjut akan dilakukan uji *log-rank*.

Berdasarkan tabel distribusi di atas didapatkan dari 69 pasien yang mendapatkan *treatment standard,* 5 (7%) di antara pasien tersebut meninggal dunia. Dari 68 pasien yang mendapatkan *treatment alternative* didapatkan 4 (5,8%) orang meninggal dunia. Terlihat bahwa proporsi pasien meninggal berdasarkan jenis *treatment* yang diambil berbeda sangat kecil. Untuk menguji hasil hipotesis jenis *treatment* yang digunakan tidak berpengaruh signifikan dengan kondisi *survival* pasien dapat dilihat dengan uji *log-ran*k. Dengan hipotesis null jenis *treatment* yang diambil tidak bermakna secara statistik, berikut adalah hasil uji *log-rank*

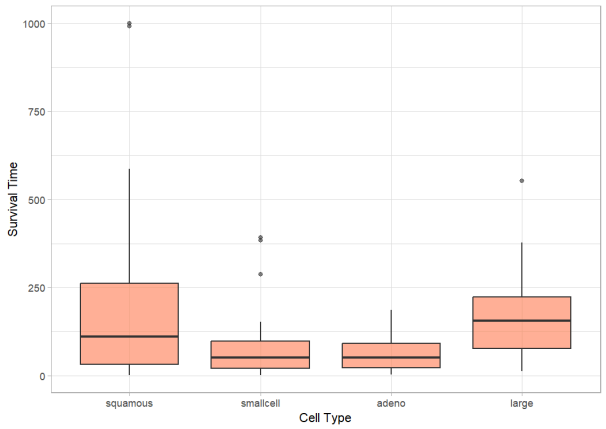
```
## Uji Log Rank
survdiff(Surv(time, status) ~ trt, data=veteran)

## Call:
## survdiff(formula = Surv(time, status) ~ trt, data = veteran)
##
##         N Observed Expected (O-E)^2/E (O-E)^2/V
## trt=1 69       64     64.5   0.00388   0.00823
## trt=2 68       64     63.5   0.00394   0.00823
##
##  Chisq= 0  on 1 degrees of freedom, p= 0.9
```

Dengan *p-value* 0.9 > 0.05 maka didapatkan hasil hipotesis null tidak ditolak atau jenis *treatment* tidak bermakna secara statistik terhadap kondisi *survival* pasien.

### *3) Survival Experience* Berdasarkan Tipe Sel

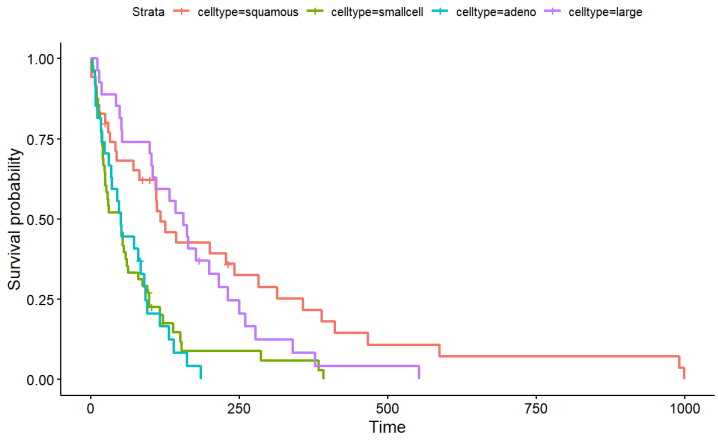
Dari 35 orang dengan tipe sel `squamous`, 4 orang di antaranya meninggal dan 31 orang lainnya bertahan hidup hingga akhir masa studi. Sementara itu, dari 48 orang dengan tipe sel `smallcell`, 3 orang di antaranya meninggal dan 45 orang lainnya bertahan hidup hingga akhir masa studi. Untuk tipe sel adeno dan large, masing-masing memiliki total 27 orang di mana 1 orang meninggal dan 26 sisanya mampu bertahan hidup hingga akhir masa studi.

```
## Tabel Distribusi
table(veteran$status, veteran$celltype)
##     squamous smallcell adeno large
##   0        4         3     1     1
##   1       31        45    26    26
```




Dari boxplot di atas, dapat dilihat bahwa tipe sel *smallcell* dan *adeno* cenderung memiliki waktu survival yang lebih singkat dengan variasi yang lebih rendah dibanding tipe sel yang lainnya. Pada *box-plot* terlihat beberapa *outlier* yang tidak dibuang untuk analisa lebih lanjut model. Untuk analisis lebih lanjut akan dilakukan uji *log-rank*.

Lebih lanjut akan dilihat grafik fungsi *survival*-nya.




Berdasarkan grafik survival probability diatas, dapat kita lihat pada hari pertama, keempat sel memiliki survival probability yang sama yaitu 100%. Kemudian, seiring berjalannya waktu, survival probability dari keempat sel terus menurun. Pada waktu ke-100, sel large terlihat memiliki survival probability yang paling besar dengan 75%. Diikuti dengan sel squamos sebesar 50%, sel adeno dan smallcell yang bernilai kisaran 25%. Kemudian, terlihat bahwa keempat sel memiliki 0% survival probability pada waktu yang berbeda beda, dengan sel adeno yang tercepat pada waktu ke-200, smallcell pada waktu ke-325, kemudian sel large pada waktu ke-550, dan squamos cell pada waktu ke-1000. Tipe sel squamosa dan smallcell terlihat memiliki tingkat survival paling rendah dibanding tipe lainnya. Hal ini selaras dengan yang ditemukan pada boxplot. Secara sekilas, squamos cell

memiliki survival probability yang lebih besar jika dibandingkan dengan sel lainnya.

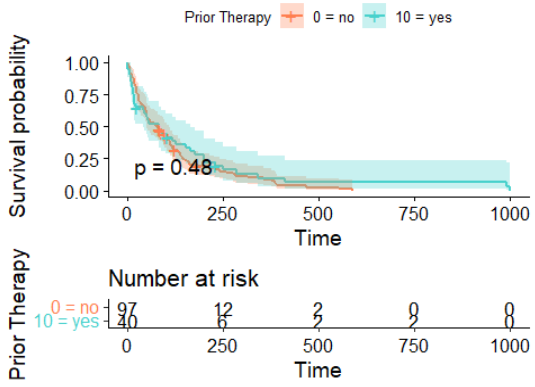
Untuk menguji apakah perbedaan fungsi survival antara keempat tipe sel cukup bermakna secara statistik, maka dilakukan uji log rank.

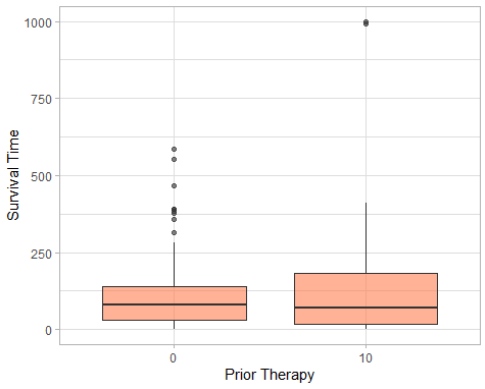
```
## Uji Log Rank
survdiff(Surv(time, status) ~ celltype, data=veteran)

## Call:
## survdiff(formula = Surv(time, status) ~ celltype, data = veteran)
##
##                       N Observed Expected (O-E)^2/E (O-E)^2/V
## celltype=squamous    35       31     47.7      5.82     10.53
## celltype=smallcell   48       45     30.1      7.37     10.20
## celltype=adeno       27       26     15.7      6.77      8.19
## celltype=large       27       26     34.5      2.12      3.02
##
##  Chisq= 25.4  on 3 degrees of freedom, p= 1e-05
```

Hasil dari uji *log-rank* menunjukkan bahwa terdapat perbedaan yang signifikan pada variabel cell type yang terdiri atas 4 kelompok sel, yaitu squamos cell, small cell, adeno cell, dan large cell. Hal ini disebabkan nilai p-value yang lebih kecil dari 0.05. Dimana p-value = 0.00001 < 0.05

**4) *Survival Experience* Berdasarkan Prior**

Secara berturut-turut berikut adalah plot Estimasi Kaplan Meier, Box-plot, dan Tabel Distribusi dari variable "prior"




```
## Tabel Distribusi
table(veteran$status, veteran$prior)
##      0 10
##   0  6  3
##   1 91 37
```

Secara sekilas pada *boxplot*, tidak terdapat perbedaan yang signifikan antara pasien yang pernah melakukan terapi dengan pasien yang tidak

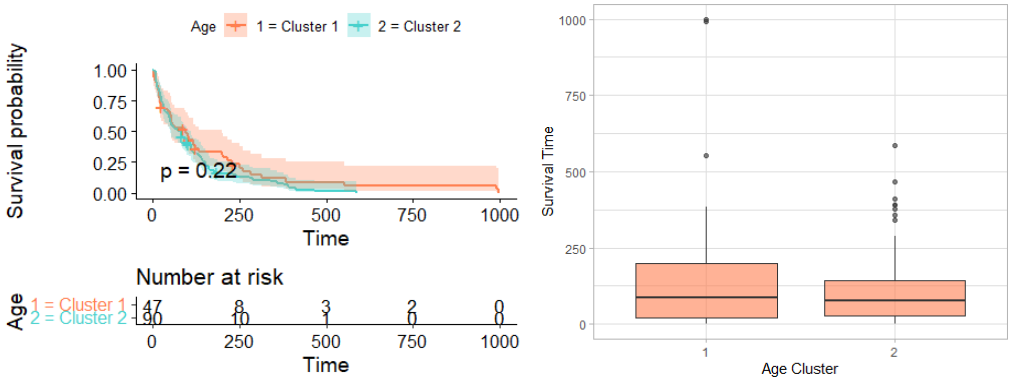
pernah melakukan terapi. Pada grafik survival, kemampuan bertahan hidup orang yang belum pernah melakukan terapi (prior=0) sedikit rendah dibanding yang pernah melakukan terapi. Hal ini terlihat pada grafik fungsi survival yang berhimpit antara kedua kelompok pasien mengindikasikan kemungkinan tidak adanya perbedaan kemampuan bertahan hidup pada pasien berdasarkan variabel "prior". Pada *box-plot* terlihat beberapa *outlier* yang tidak dibuang untuk analisa lebih lanjut model. Untuk analisis lebih lanjut akan dilakukan uji *log-rank*.

```
## Uji Log Rank
survdiff(Surv(time, status) ~ prior, data=veteran)

## Call:
## survdiff(formula = Surv(time, status) ~ prior, data = veteran)
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## prior=0  97       91     87.4     0.150     0.501
## prior=10 40       37     40.6     0.323     0.501
##
##  Chisq= 0.5  on 1 degrees of freedom, p= 0.5
```

Hasil uji log-rank menunjukkan bahwa tidak terdapat perbedaan yang signifikan pada variabel prior, yang terdiri atas pasien yang pernah melakukan terapi dengan pasien yang tidak pernah melakukan terapi. Hal ini disebabkan oleh nilai p-value yang lebih besar dari 0.05, dimana p-value = 0.5 > 0.05.

**5) *Survival Experience* Berdasarkan *Age Cluster***

Secara berturut-turut berikut adalah plot Estimasi Kaplan Meier, Box-plot, dan Tabel Distribusi dari variable "age_cluster"




```
## Tabel Distribusi
table(veteran$status, veteran$age_cluster)

##
##      1  2
```

```
##   0  5  4
##   1 42 83
```

Jika melihat pada *box-plot* dan grafik di atas, terlihat tidak ada perbedaan yang mencolok antara waktu *survival* antar *age cluster*. Terlihat beberapa nilai *outlier* pada variabel ini yang tidak dibuang karena ditakutkan akan mempengaruhi hasil dari analisis selanjutnya. Untuk analisis lebih lanjut akan dilakukan uji *log-rank*.

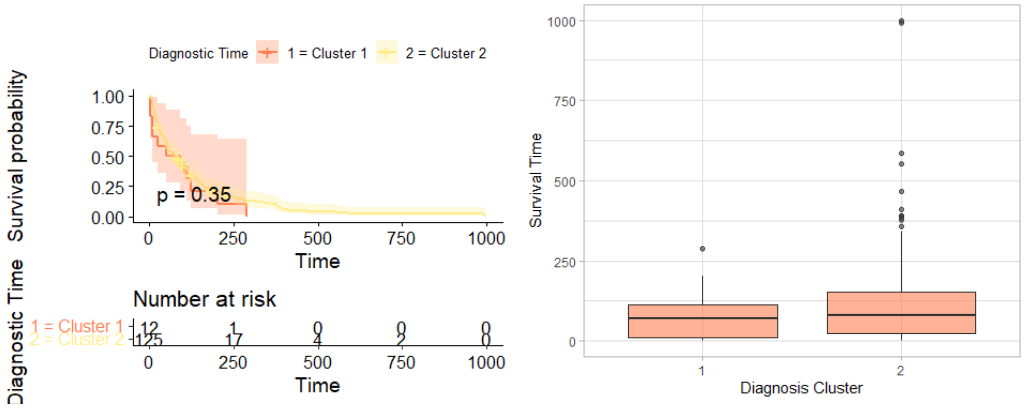
Berdasarkan tabel distribusi di atas didapatkan dari 47 pasien pada *age cluster*$_1$ , 5 (10%) di antara pasien tersebut meninggal dunia. Dari 87 pasien pada *age cluster*$_2$ didapatkan 4 (4,59%) orang meninggal dunia. Terlihat bahwa proporsi pasien meninggal pada *age cluster*$_1$ 2 kali lipat lebih besar dibandingkan *age cluster*$_2$. Dengan hipotesis null *age cluster* tidak bermakna signifikan terhadap kondisi *survival* pasien akan dilakukan uji *log-rank* berikut.

```
## Uji Log Rank
survdiff(Surv(time, status) ~ age_cluster, data=veteran)

## Call:
## survdiff(formula = Surv(time, status) ~ age_cluster, data = veteran)
##
##                 N Observed Expected (O-E)^2/E (O-E)^2/V
## age_cluster=1 47       42     48.6     0.905      1.54
## age_cluster=2 90       86     79.4     0.555      1.54
##
##  Chisq= 1.5  on 1 degrees of freedom, p= 0.2
```

Dengan *p-value* 0.2 > 0.05 maka didapatkan hasil hipotesis null tidak ditolak atau *age cluster* tidak bermakna secara statistik terhadap kondisi *survival* pasien.

### 6) *Survival Experience* Berdasarkan *Diagtime Cluster*

Secara berturut-turut berikut adalah plot Estimasi Kaplan Meier, Box-plot, dan Tabel Distribusi dari variable "diagtime_cluster"




```
## Tabel Distribusi
table(veteran$status, veteran$diagtime_cluster)
```

```
##
##      1   2
##   0  1   8
##   1 11 117
```

Jika melihat pada *box-plot* dan grafik di atas, terlihat tidak ada perbedaan yang mencolok antara waktu *survival* antar *diagtime*. Terlihat beberapa nilai *outlier* pada variabel ini yang tidak dibuang karena ditakutkan akan mempengaruhi hasil dari analisis selanjutnya. Untuk analisis lebih lanjut akan dilakukan uji *log-rank*.

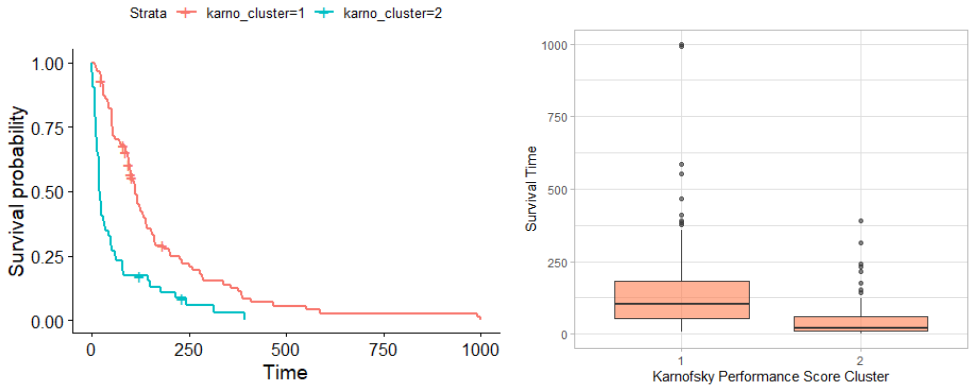
Berdasarkan tabel distribusi di atas didapatkan dari 12 pasien pada *diag cluster*$_1$ , 1 (8,3%) di antara pasien tersebut meninggal dunia. Dari 125 pasien pada *diag cluster*$_2$ didapatkan 8 (6,4%) orang meninggal dunia. Terlihat bahwa proporsi pasien meninggal pada kedua *diagtime* tidaklah berbeda jauh. Dengan hipotesis null *age cluster* tidak bermakna signifikan terhadap kondisi *survival* pasien akan dilakukan uji *log-rank* berikut.

```
## Uji Log Rank
survdiff(Surv(time, status) ~ diagtime_cluster, data=veteran)

## Call:
## survdiff(formula = Surv(time, status) ~ diagtime_cluster, data =
veteran)
##
##                       N Observed Expected (O-E)^2/E (O-E)^2/V
## diagtime_cluster=1  12       11     8.41    0.8005     0.872
## diagtime_cluster=2 125      117   119.59    0.0563     0.872
##
##  Chisq= 0.9  on 1 degrees of freedom, p= 0.4
```

Dengan *p-value* 0.4> 0.05 maka didapatkan hasil hipotesis null tidak ditolak atau *diagtime cluster* tidak bermakna secara statistik terhadap kondisi *survival* pasien.

### 7) *Survival Experience* Berdasarkan *Karnofsky Score*

Skor karnofsky adalah salah satu dari sistem penilaian status performa yang digunakan pada pasien kanker, hasil pengukuran skor karnofsky berkaitan erat dengan kualitas hidup dan keadaan fungsional fisik pasien [5]. Semakin besar skor karnofsky (maksimal 100), semakin sehat keadaaan fungsional fisik pasien. Secara berturut-turut berikut adalah plot Estimasi Kaplan Meier, Box-plot, dan Tabel Distribusi dari variable “karno_cluster”.



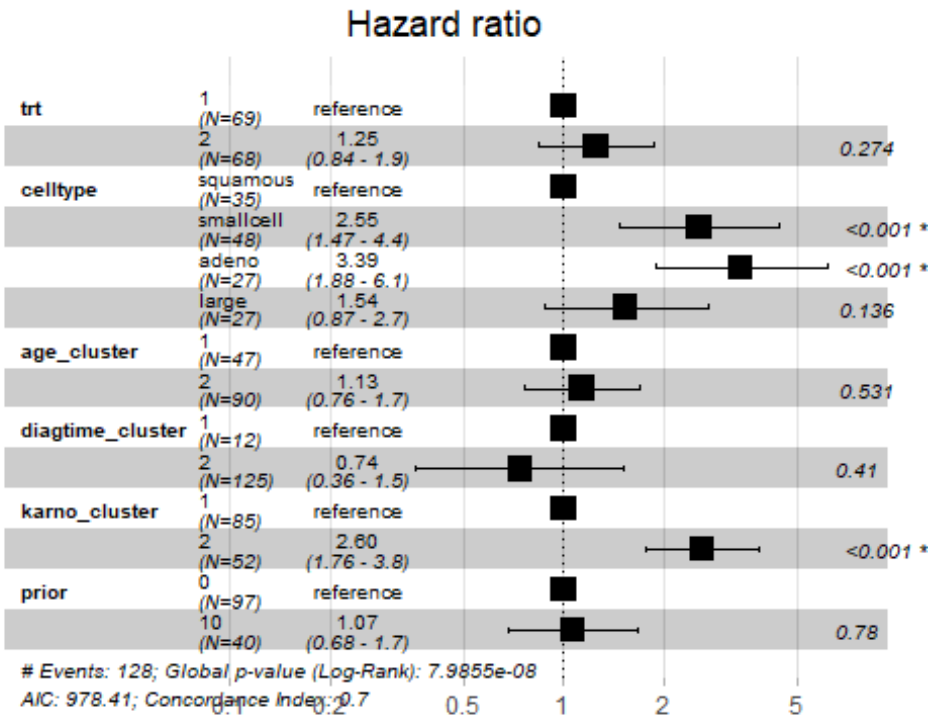

```
## Tabel Distribusi
table(veteran$status, veteran$karno_cluster)

##
##      1  2
##   0  7  2
##   1 78 50
```

Berdasarkan boxplot, terdapat perbedaan yang cukup jelas antara kedua cluster. Terlihat beberapa nilai *outlier* pada variabel ini yang tidak dibuang karena ditakutkan akan mempengaruhi hasil dari analisis selanjutnya. Berdasarkan grafik fungsi survival dan box plot kedua kelompok pasien ini terlihat cukup konsisten berbeda. Kemampuan bertahan hidup untuk untuk cluster 2 menurun secara drastis seiring berjalannya waktu. Sedangkan, cluster 1 fungsi survivenya menurun tetapi tidak sedrastis cluster 2. Oleh karena itu, dapat diambil kesimpulan bahwa terdapat perbedaan yang signifikan antara karno cluster 1 dan karno cluster 2. Untuk analisis lebih lanjut akan dilakukan uji *log-rank*.

```
## Uji Log Rank
survdiff(Surv(time, status) ~ karno_cluster, data=veteran)

## Call:
## survdiff(formula = Surv(time, status) ~ karno_cluster, data =
veteran)
##
##                  N Observed Expected (O-E)^2/E (O-E)^2/V
## karno_cluster=1 85       78    101.7      5.52      28.2
## karno_cluster=2 52       50     26.3     21.36      28.2
##
##  Chisq= 28.2  on 1 degrees of freedom, p= 1e-07
```

Hasil uji log-rank mendukung bahwa terdapat perbedaan yang signifikan pada karno cluster yang terdiri atas karno cluster 1 dan karno cluster 2. Hal ini ditunjukkan dari nilai p-value yang jauh lebih kecil dari 0.05. Dimana p-value = 1e^-7 < 0.05.

## 8) Stratified Plot with ggforest



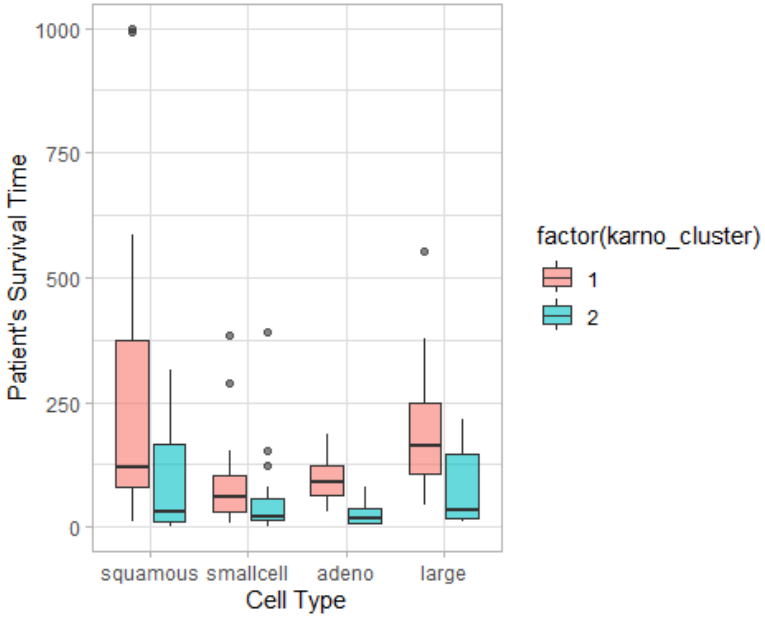

Berikut adalah visualisasi dengan *ggforest* yang merangkum hazard ratio dari semua variabel yang akan dipakai untuk membentuk model Cox PH selanjutnya.

Analisis *hazard ratio* menggunakan *ggforest* pada R gambar di atas menunjukkan bahwa variabel "celltype" (tipe sel) dan "karno_cluster" (Karnofsky score) memiliki hubungan yang signifikan dengan risiko kematian pasien.

- celltype: Individu dengan tipe sel "smallcell" memiliki risiko 2,55 kali lebih tinggi daripada individu dengan tipe sel "squamous".. Hal ini juga berlaku untuk individu dengan tipe sel "adeno" yang memiliki risiko 3,39 kali lebih tinggi dibandingkan dengan individu dengan tipe sel "squamous".
- karno_cluster: Individu dalam cluster 2 memiliki risiko 2,60 kali lebih tinggi dibandingkan dengan cluster 1.

Variabel lain seperti tidak menunjukkan hubungan yang signifikan secara statistik dengan risiko kematian pasien.

**9) Boxplot variabel "celltype" dan "karno_cluster"**




Berdasarkan boxplot diatas, dapat dilihat bahwa pasien yang memiliki celltype squamos dengan faktor karno_cluster 1 memiliki rata-rata survive time yang lebih panjang yakni dengan waktu terpendek selama 10 dan waktu terpanjang selama 600. Sementara itu, pasien yang memiliki smallcell dan adeno cell dengan faktor karno cluster 2 memiliki waktu survive time rata-rata terendah yakni dengan waktu terendah selama 0 hari dan waktu terlama 60 hari.

**10) Kesimpulan**

Kesimpulan yang dapat kita ambil adalah variabel celltype dan karno_cluster merupakan variabel penjelas yang potensial untuk fungsi survival dari kejadian pasien yang tidak survive. Hal ini dikarenakan berdasarkan grafik fungsi survival probability dan uji log-rank didapatkan kesimpulan bahwa terdapat perbedaan yang signifikan pada keempat kelompok sel pada variabel celltype dan kedua kelompok pada variabel karno_cluster. Artinya, perbedaan jenis sel dan perbedaan kemampuan pasien dalam melakukan sebuah aktivitas akan mempengaruhi kemungkinan pasien untuk hidup

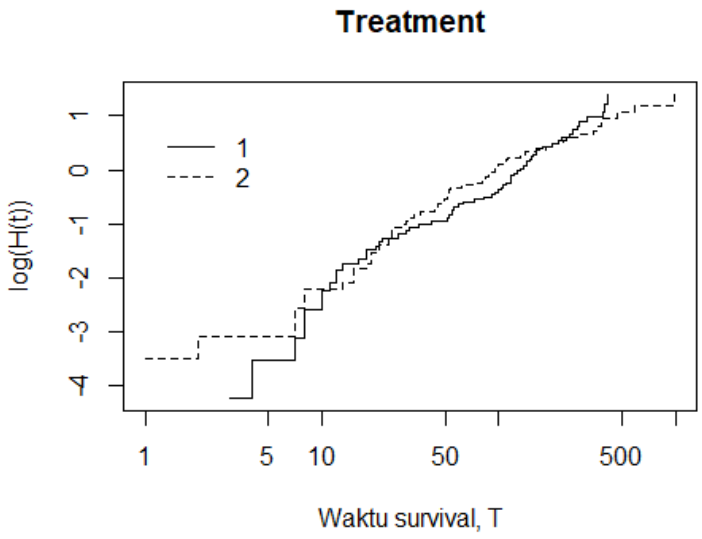
# BAB III MODEL ANALYSIS

## A. Uji Cox PH secara Grafis

Pada analisis survival, nilai hazard antar grup di titik tertentu memiliki nilai yang berbeda-beda dan umumnya nilai hazard tidaklah konstan. Namun, perlu diperhatikan bahwa rasio hazard antar grup diasumsikan konstan terhadap waktu. Atas dasar tersebut maka digunakanlah uji asumsi PH dengan metode grafis yang menggambarkan plot logaritma dan fungsi kumulatif hazard dari beberapa kelompok.

### 1) Uji asumsi PH dengan metode grafis pada variabel "trt"

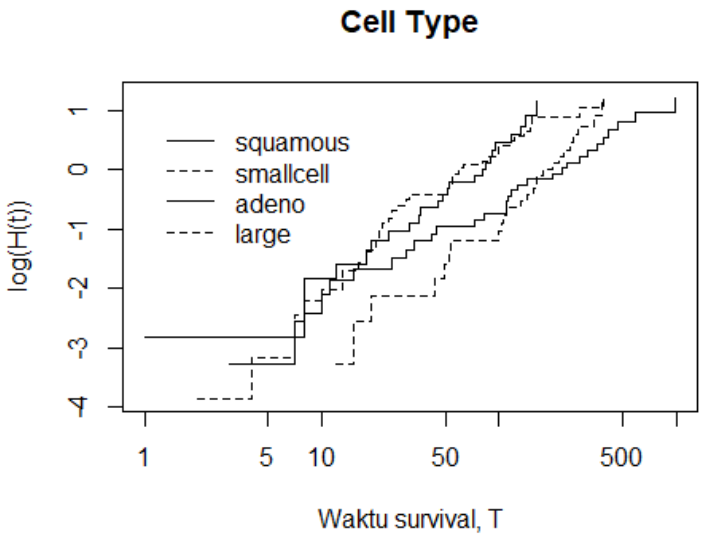
Berikut adalah plot logaritma dan fungsi kumulatif hazard untuk dua kelompok variabel "trt" (*treatment*).




Berdasarkan plot di atas terlihat ketidapararelan kurva untuk variabel "trt" dimana fungsi $log(H(t))$ yang berhimpit pada waktu yang relatif lama.

### 2) Uji asumsi PH dengan metode grafis pada variabel "celltype"

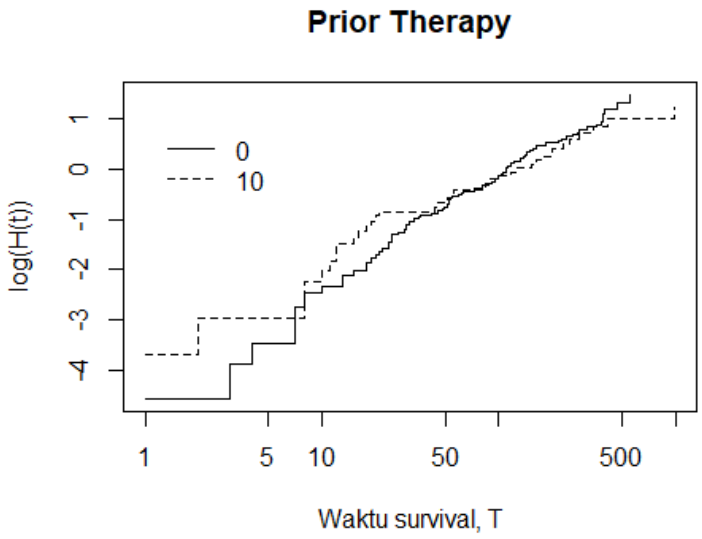
Berikut adalah plot logaritma dan fungsi kumulatif hazard untuk keempat kelompok variabel "celltype" (tipe sel)




Berdasarkan plot di atas diduga kondisi survival tiap pasien dengan tipe cell yang berbeda mengalami perubahan yang berbeda-beda seiring berjalannya waktu. Terlihat pada tipe cell "smallcell" memiliki kenaikan fungsi survival yang sangat cepat relatif terhadap tipe lain yang mengindikasikan efek dari tipe lain terhadap survival. Pada tipe cell "squamous" dan "large" terlihat tidak berbeda secara signifikan yang mengindikasikan tidak ada efek yang berarti dari tipe tersebut terhadap fungsi survival.

**3) Uji asumsi PH dengan metode grafis pada variabel "prior"**

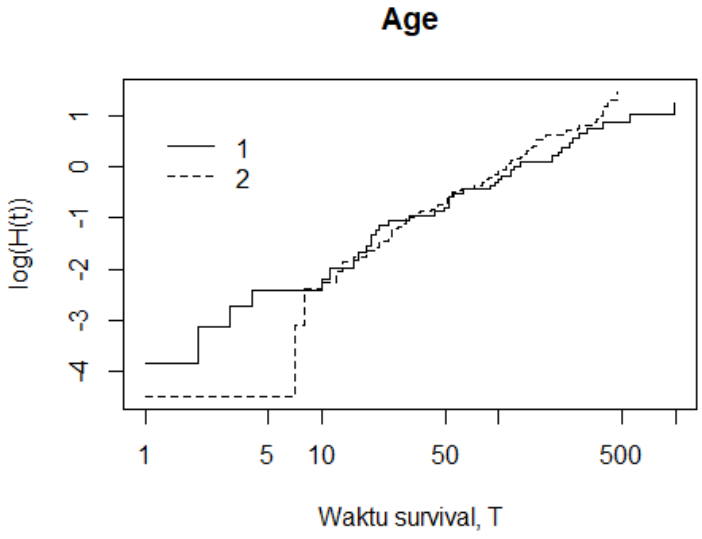
Berikut adalah plot logaritma dan fungsi kumulatif hazard untuk kedua kelompok variabel "prior"




Berdasarkan plot di atas terlihat ketidakpararelan kurva untuk variabel "prior" dimana fungsi $log(H(t))$ yang berhimpit pada awal, tengah, dan akhir waktu yang relatif lama.

**4) Uji asumsi PH dengan metode grafis pada variabel "age_cluster"**

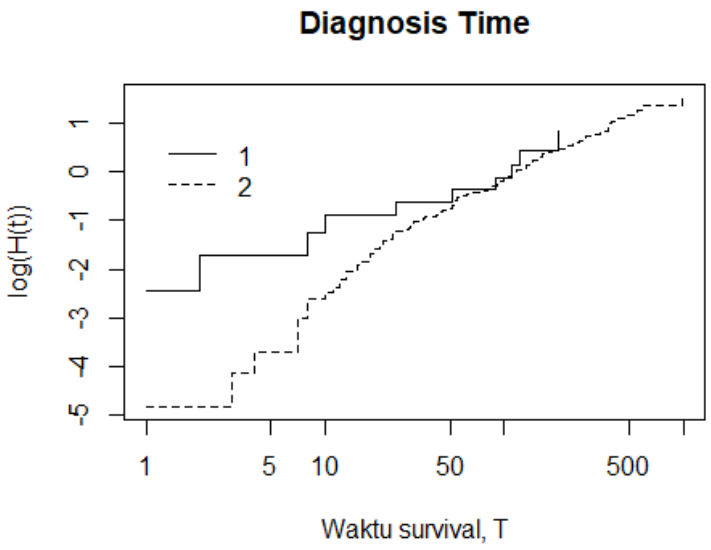
Berikut adalah plot logaritma dan fungsi kumulatif hazard untuk kedua kelompok variabel "age_cluster"




Berdasarkan plot di atas terlihat ketidakpararelan kurva untuk variabel "age" dimana fungsi $log(H(t))$ yang berhimpit pada waktu yang relatif lama.

**5) Uji asumsi PH dengan metode grafis pada variabel "diagtime_cluster"**

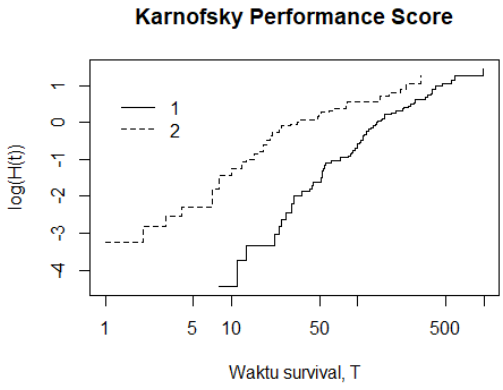
Berikut adalah plot logaritma dan fungsi kumulatif hazard untuk kedua kelompok variabel "diagtime_cluster"




Berdasarkan plot di atas terlihat ketidaksejajaran kurva untuk variabel "diagnosis time" dimana fungsi $log(H(t))$ berbeda dengan diagnosis pada kluster 2 yang relatif linear terhadap waktu.

**6) Uji asumsi PH dengan metode grafis pada variabel "karno_cluster"**

Berikut adalah plot logaritma dan fungsi kumulatif hazard untuk kedua kelompok variabel "karno_cluster"




Pada plot di atas cukup terlihat kesejajaran fungsi $log(H(t))$ namun tidak sempurna.

## B. Model yang Diajukan

### 1) Model 1 (Menggunakan Semua Kovariat)

Berikut hasil uji Cox-PH dari model 1 yang menggunakan seluruh kovariat dengan bantuan R

```
# Model 1: Use all covariate ----

model1 <- coxph(Surv(time, status) ~ trt + celltype + age_cluster +
diagtime_cluster + karno_cluster + prior, data = veteran,
                method='breslow')
summary(model1)

## Call:
## coxph(formula = Surv(time, status) ~ trt + celltype + age_cluster +
##     diagtime_cluster + karno_cluster + prior, data = veteran,
##     method = "breslow")
##
##   n= 137, number of events= 128
##                       coef exp(coef) se(coef)      z Pr(>|z|)
## trt2                0.2205    1.2466   0.2048  1.077 0.281671
## celltypesmallcell1  0.9291    2.5323   0.2798  3.321 0.000897 ***
## celltypeadeno       1.2130    3.3636   0.3009  4.032 5.54e-05 ***
## celltypelarge       0.4280    1.5342   0.2880  1.486 0.137225
## age_cluster2        0.1252    1.1333   0.2015  0.621 0.534607
## diagtime_cluster2 -0.3002    0.7406   0.3662 -0.820 0.412296
## karno_cluster2      0.9510    2.5882   0.1985  4.791 1.66e-06 ***
## prior10             0.0625    1.0645   0.2287  0.273 0.784618
```

Dari hasil R di atas didapatkan model 1 regresi Cox-PH dengan persamaan

$$h(t) = h_0(t)exp(\beta_1 trt_2 + \beta_2 celltypesmallcell + \beta_3 celltypeadeno + \beta_4 celltypelarge + \beta_5 age\_cluster_2 + \beta_6 diagtime\_cluster_2 + \beta_7 karno\_cluster_2 + \beta_8 prior_{10})$$

Model ini cukup signifikan dalam menjelaskan kondisi pasien dengan *cell type small* dan *adeno* serta karno tipe 2.

```
# Asumsi Cox PH
cox.zph(model1)

##                     chisq df       p
## trt                0.0394  1 0.84262
## celltype          12.2896  3 0.00645
## age_cluster        1.4788  1 0.22396
## diagtime_cluster   0.1451  1 0.70325
```

```
## karno_cluster    15.4209  1 8.6e-05
## prior             2.1574  1 0.14189
## GLOBAL           29.5373  8 0.00026
```

Dengan menguji asumsi Cox PH pada model 1 didapatkan variabel "karno_cluster" memiliki *p-value* < 0,05 sehingga asumsi *proportional hazard* tidak terpenuhi pada model ini. Secara global didapatkan juga berdasarkan *p-value* menolak asumsi PH.

**2) Model 2 (Menggunakan Semua Kovariat dan *Backward* AIC)**

Dikarena pada model 1 masih terdapat variabel yang tidak memenuhi asumsi *proportional hazard*, maka akan digunakan bantuan *backward* AIC pada model 2. Berikut hasil uji Cox-PH dari model 2 yang menggunakan seluruh kovariat dan *backward* AIC dengan bantuan R. Perlu diperhatikan bahwa mekanisme *backward* AIC adalah dengan menghapuskan variabel yang memiliki nilai AIC terkecil pada model 2. Setelah dilakukan *backward* AIC, digunakan 2 variabel yaitu "celltype" dan "karno_cluster"

```
# Model 2: Use all covariate, Backward AIC ----
model2 <- stepAIC(object=model1, data=veteran_data,
direction="backward")

summary(model2)

## Call:
## coxph(formula = Surv(time, status) ~ celltype +
karno_cluster,
##     data = veteran, method = "breslow")
##
##   n= 137, number of events= 128
##
##                      coef exp(coef) se(coef)     z Pr(>|z|)
## celltypesmallcell 0.8643    2.3734   0.2533 3.413 0.000643 ***
## celltypeadeno     1.1916    3.2924   0.2932 4.064 4.83e-05 ***
## celltypelarge     0.3278    1.3879   0.2775 1.181 0.237475
## karno_cluster2    0.9032    2.4674   0.1937 4.662 3.13e-06 ***
```

Dari hasil R di atas didapatkan model 1 regresi Cox-PH dengan persamaan $h(t) = h_0(t)exp(\beta_1 celltypesmallcell + \beta_2 celltypeadeno + \beta_3 celltypelarge + \beta_4 karno\_cluster_2)$.

Model ini cukup signifikan dalam menjelaskan kondisi pasien dengan *cell type small* dan *peadeno* serta karno tipe 2.

```
# Asumsi Cox PH
cox.zph(model2)

##                chisq df       p
## celltype        11.7  3  0.0083
## karno_cluster   16.8  1 4.1e-05
## GLOBAL          25.0  4 5.1e-05
```

Dengan menguji asumsi Cox PH pada model 2 didapatkan semua variabel tidak memenuhi asumsi PH karena nilai *p-value* < 0,05 untuk semua variabel.

**3) Model 3 (Menggunakan Semua Kovariat dan Strata "karno_cluster")**

Diketahui sebelumnya pada model 1 dan model 2 variabel "karno_cluster" tidak memenuhi asumsi PH, maka dari itu variabel ini akan dikeluarkan dari model regresi. Namun, karena pada variabel "karno_cluster" memiliki fungsi survival yang berbeda untuk 2 tipe karno yang berbeda maka perlu modifikasi dengan memperhitungkan tetap fungsi hazard pada variabel "karno_cluster" yaitu dengan stratifikasi model. Berikut hasil uji Cox-PH dari model 3 yang menggunakan seluruh kovariat dan strata pada variabel "karno_cluster".

```
model3 <- coxph(Surv(time, status) ~ trt + celltype + age_cluster +
diagtime_cluster + strata(karno_cluster) + prior, data = veteran,
                    method='breslow')
summary(model3)

## Call:
## coxph(formula = Surv(time, status) ~ trt + celltype + age_cluster +
##     diagtime_cluster + strata(karno_cluster) + prior, data = veteran,
##     method = "breslow")
##
##   n= 137, number of events= 128
##
##                         coef exp(coef)  se(coef)      z Pr(>|z|)
## trt2                0.099693  1.104832  0.202700  0.492 0.622841
## celltypesmallcell1  0.912214  2.489829  0.281987  3.235 0.001217 **
## celltypeadeno       1.097507  2.996687  0.303088  3.621 0.000293 ***
## celltypelarge       0.336083  1.399455  0.288831  1.164 0.244588
## age_cluster2        0.121540  1.129235  0.201878  0.602 0.547142
## diagtime_cluster2 -0.180644  0.834733  0.363300 -0.497 0.619026
## prior10             0.009038  1.009079  0.225976  0.040 0.968097
```

Dari hasil R di atas didapatkan model 1 regresi Cox-PH dengan persamaan

$$h_s(t) = h_{0,s}(t)exp(\beta_1 trt_2 + \beta_2 celltypesmallcell + \beta_3 celltypeadeno + \beta_4 celltypelarge + \beta_5 age\_cluster_2 + \beta_6 diagtime\_cluster_2 + \beta_7 prior_{10}),\ s = 1, 2$$

Model ini cukup signifikan dalam menjelaskan kondisi pasien dengan *cell type small* dan *peadeno*.

```
# Asumsi Cox PH
cox.zph(model3)
```

```
##                chisq df       p
## trt            0.586  1 0.444
## celltype       7.954  3 0.047
## age_cluster    2.197  1 0.138
## diagtime_cluster  0.727  1 0.394
## prior          2.890  1 0.089
## GLOBAL        12.352  7 0.090
```

Dengan menguji asumsi Cox PH pada model 3 didapatkan hanya variabel "celltype" yang tidak memenuhi asumsi *proportional hazard* dengan *p-value* < 0,05. Secara global, model 3 memenuhi asumsi PH.

**4) Model 3.5 (Menggunakan Semua Kovariat, Strata "karno_cluster", Menghilangkan "prior")**

Pada model 3, dapat dilihat bahwa *p-value* uji asumsi Cox PH untuk variabel "`celltype`" sebesar `0.047`. Nilai ini nyaris mendekati batas 0.05. Oleh karena itu, pada bagian ini, kami mencoba untuk mengeluarkan 1 variabel dengan *p-value* uji parameter paling besar, yakni variabel "prior" dengan *p-value* uji parameter sebesar 0.968, dan melihat apakah terdapat pemenuhan uji asumsi Cox PH pada variabel "celltype". Berikut hasil uji Cox-PH dari model 2 yang menggunakan seluruh kovariat, strata pada "karno_cluster", dan menghilangkan variabel "prior".

```
model3.5 <- coxph(Surv(time, status) ~ trt + celltype + age_cluster +
diagtime_cluster + strata(karno_cluster), data = veteran,
                  method='breslow')
summary(model3.5)
## Call:
## coxph(formula = Surv(time, status) ~ trt + celltype + age_cluster +
##     diagtime_cluster + strata(karno_cluster), data = veteran,
##     method = "breslow")
```

```
##
##   n= 137, number of events= 128
##
##                       coef exp(coef) se(coef)      z Pr(>|z|)
## trt2                0.1000    1.1052   0.2026  0.494 0.621430
## celltypesmallcell   0.9107    2.4861   0.2794  3.260 0.001116 **
## celltypeadeno       1.0955    2.9906   0.2988  3.667 0.000246 ***
## celltypelarge       0.3363    1.3998   0.2888  1.165 0.244121
## age_cluster2        0.1207    1.1283   0.2009  0.601 0.547845
## diagtime_cluster2 -0.1862    0.8301   0.3359 -0.554 0.579376
```

Dari hasil R di atas didapatkan model 1 regresi Cox-PH dengan persamaan

$$h_s(t) = h_{0,s}(t)exp(\beta_1 trt_2 + \beta_2 celltypesmallcell + \beta_3 celltypeadeno + \beta_4 celltypelarge + \beta_5 age\_cluster_2 + \beta_6 diagtime\_cluster_2)$$

$$, s = 1, 2.$$

Model ini cukup signifikan dalam menjelaskan kondisi pasien dengan *cell type small* dan *peadeno*.

```
# Asumsi Cox PH
cox.zph(model3.5)
```

```
##                chisq df       p
## trt                0.579  1 0.447
## celltype           7.837  3 0.050
## age_cluster        2.197  1 0.138
## diagtime_cluster   0.722  1 0.395
## GLOBAL            11.066  6 0.086
```

Dengan menguji asumsi Cox PH pada model 3.5 didapatkan semua variabel dan model 3.5 memenuhi asumsi *proportional hazard* terlihat dari *p-value* > 0,05. Secara global, asumsi PH terpenuhi pada model 3.5.

**5) Model 4 (Menggunakan Semua Kovariat dan Strata "celltype")**

Diketahui sebelumnya pada model 2 variabel "celltype" dan "karno_cluster" tidak memenuhi asumsi PH, maka dari itu variabel ini akan dikeluarkan dari model regresi. Namun, karena pada variabel "celltype" memiliki fungsi survival yang berbeda untuk 4 tipe sell yang berbeda maka perlu modifikasi dengan memperhitungkan tetap fungsi hazard pada variabel "celltype" yaitu dengan stratifikasi model. Berikut hasil uji Cox-PH dari model 4 yang menggunakan seluruh kovariat dan strata pada variabel "celltype".

```
model4 <- coxph(Surv(time, status) ~ trt + strata(celltype) +
age_cluster + diagtime_cluster + karno_cluster + prior, data = veteran,
                      method='breslow')
summary(model4)
## Call:
## coxph(formula = Surv(time, status) ~ trt + strata(celltype) +
##       age_cluster + diagtime_cluster + karno_cluster + prior, data =
veteran,
##     method = "breslow")
##
##   n= 137, number of events= 128
##
##                       coef exp(coef) se(coef)      z Pr(>|z|)
## trt2                0.1672    1.1820   0.2072  0.807    0.420
## age_cluster2        0.1259    1.1342   0.2069  0.609    0.543
## diagtime_cluster2 -0.3203    0.7259   0.3658 -0.875    0.381
## karno_cluster2      1.0567    2.8769   0.2078  5.084 3.69e-07 ***
## prior10             0.1162    1.1233   0.2311  0.503    0.615
```

Dari hasil R di atas didapatkan model 4 regresi Cox-PH dengan persamaan

$$h_s(t) = h_{0,s}(t)exp(\beta_1 trt_2 + \beta_2 age\_cluster_2 + \beta_3 diagtime\_cluster_2 + \beta_4 karno\_cluster_2 + \beta_5 prior_{10}),\ s = 1,2,3,4$$

Model ini signifikan dalam menjelaskan kondisi pasien dengan karno tipe 2.

```
# Asumsi Cox PH
cox.zph(model4)

##                    chisq df       p
## trt                0.675  1 0.41141
## age_cluster        1.388  1 0.23881
## diagtime_cluster   0.416  1 0.51890
## karno_cluster     12.039  1 0.00052
## prior              1.324  1 0.24981
## GLOBAL            17.241  5 0.00406
```

Dengan menguji asumsi Cox-PH pada model 4 didapatkan variabel "karno_cluster" dan model 4 secara global tidak memenuhi asumsi *proportional hazard* karena *p-value* < 0,05.

**6) Model 5 (Menggunakan Semua Kovariat dan Strata "celltype" serta "karno_cluster")**

Pada model 3 dan 4 telah dilakukan uji regresi Cox-Stratifikasi pada variabel "celltype" dan "karno_cluster" dengan hasil tidak memenuhi asumsi PH baik dengan uji regresi Cox-PH biasa dan Cox-Stratifikasi. Ditambah berdasarkan

model 2 didapatkan pada kedua variabel tersebut tidak memenuhi asumsi PH, maka dari itu akan diuji apakah dengan menstratifikasi kedua variabel yang tidak memenuhi asumsi PH tersebut dihasilkan model yang bagus. Berikut hasil uji Cox-PH dari model 5 yang menggunakan seluruh kovariat dan strata pada variabel "celltype" dan "karno_cluster".

```
model5 <- coxph(Surv(time, status) ~ trt + strata(celltype) +
age_cluster + diagtime_cluster + strata(karno_cluster) + prior, data =
veteran,
                    method='breslow')
summary(model5)
## Call:
## coxph(formula = Surv(time, status) ~ trt + strata(celltype) +
##     age_cluster + diagtime_cluster + strata(karno_cluster) +
##     prior, data = veteran, method = "breslow")
##
##   n= 137, number of events= 128
##
##                       coef exp(coef) se(coef)      z Pr(>|z|)
## trt2               0.07766   1.08075  0.20748  0.374    0.708
## age_cluster2       0.16587   1.18042  0.21269  0.780    0.435
## diagtime_cluster2 -0.13952   0.86978  0.36799 -0.379    0.705
## prior10            0.09045   1.09467  0.22988  0.393    0.694
```

Dari hasil R di atas didapatkan model 5 regresi Cox-PH dengan persamaan

$$h_s(t) = h_{0,s}(t)exp(\beta_1 trt_2 + \beta_2 age\_cluster_2 + \beta_3 diagtime\_cluster_2 + \beta_4 prior_{10}),\ s = 1,2,3,4,5,6,7,8$$

Tidak ada variabel yang signifikan pada model ini.

```
# Asumsi Cox PH
cox.zph(model5)

##                    chisq df     p
## trt                0.5129  1 0.474
## age_cluster        3.3183  1 0.069
## diagtime_cluster 0.0297  1 0.863
## prior              1.3688  1 0.242
## GLOBAL             5.2824  4 0.260
```

Dengan menguji asumsi Cox PH pada model 5 didapatkan seluruh variabel dan model 5 memenuhi asumsi *proportional hazard* dengan *p-value* > 0,05.

**7) Model 6 (Menggunakan Semua Kovariat, Strata "karno_cluster", *Backward* AIC)**

Pada model 5 telah dilakukan uji regresi Cox-Stratifikasi dengan 2 strata variabel sekaligus, akan dibuat model 6 sebagai pertimbangan atas model 5. Model 6 merupakan model yang dirancang untuk menyelesaikan masalah pada model 3 yang memiliki pelanggaran asumsi pada variabel "celltype". Berikut hasil uji Cox-PH dari model 6 yang menggunakan seluruh kovariat, strata pada variabel "karno_cluster", dan *backward* AIC. Dengan mekanisme *backward* AIC, variabel "prior", "trt", "diagtime_cluster", dan "age_cluster" secara berurutan memiliki nilai AIC yang besar sehingga perlu dihilangkan. Hasil dari *backward* AIC ini menyisakan variabel "celltype" dan strata pada "karno_cluster".

```
model6_draft <- coxph(Surv(time, status) ~ trt + celltype + age_cluster
+ diagtime_cluster + strata(karno_cluster) + prior, data = veteran,
                method='breslow')
model6      <-      stepAIC(object=model6_draft,      data=veteran_data,
direction="backward")
summary(model6)

## Call:
##    coxph(formula    =    Surv(time,    status)    ~    celltype    +
strata(karno_cluster),
##     data = veteran, method = "breslow")
##
##   n= 137, number of events= 128
##
##                       coef exp(coef) se(coef)     z Pr(>|z|)
## celltypesmallcell 0.8920    2.4400   0.2604 3.425 0.000615 ***
## celltypeadeno     1.0953    2.9902   0.2963 3.697 0.000218 ***
## celltypelarge     0.2833    1.3275   0.2796 1.013 0.310876
```

Dari hasil R di atas didapatkan model 6 regresi Cox-PH dengan persamaan

$h_s(t) = h_{0,s}(t)exp(\beta_1 celltypesmallcell + \beta_2 celltypeadeno + \beta_3 celltypelarge)$

, $s = 1, 2$

Model ini cukup signifikan dalam menjelaskan kondisi pasien dengan *cell type small* dan *peadeno*.

```
# Asumsi Cox PH
cox.zph(model6)
##          chisq df    p
## celltype  7.42  3 0.06
## GLOBAL    7.42  3 0.06
```

Dengan menguji asumsi Cox PH pada model 6 didapatkan seluruh variabel dan model 6 secara global memenuhi asumsi *proportional hazard* dengan *p-value* > 0,05.

**8) Model 7 (Menggunakan Semua Kovariat, Strata "celltype", Backward AIC)**

Pada model 5 telah dilakukan uji regresi Cox-Stratifikasi dengan 2 strata variabel sekaligus, akan dibuat model 7 sebagai pertimbangan atas model 5 dan 6. Model 7 merupakan model yang dirancang untuk menyelesaikan masalah pada model 4 yang memiliki pelanggaran asumsi pada variabel "karno_cluster". Berikut hasil uji Cox-PH dari model 7 yang menggunakan seluruh kovariat, strata pada variabel "celltype", dan *backward* AIC. Dengan mekanisme *backward* AIC, variabel "prior", "trt", "diagtime_cluster", dan "age_cluster" secara berurutan memiliki nilai AIC yang besar sehingga perlu dihilangkan. Hasil dari *backward* AIC ini menyisakan variabel "celltype" dan strata pada "karno_cluster".

```
model7_draft <- coxph(Surv(time, status) ~ trt + strata(celltype) +
age_cluster + diagtime_cluster + karno_cluster + prior, data = veteran,
                      method='breslow')
model7      <-      stepAIC(object=model7_draft,      data=veteran_data,
direction="backward")
summary(model7)
## Call:
##   coxph(formula   =   Surv(time,   status)   ~   strata(celltype)   +
karno_cluster,
##     data = veteran, method = "breslow")
##
##   n= 137, number of events= 128
##
##                   coef exp(coef) se(coef)     z Pr(>|z|)
## karno_cluster2 1.0082    2.7406   0.2043 4.935 8.03e-07 ***
```

Dari hasil R di atas didapatkan model 7 regresi Cox-PH dengan persamaan $h_s(t) = h_{0,s}(t)exp(\beta_1 karno\_cluster_2)$, $s = 1, 2, 3, 4$. Model ini cukup signifikan dalam menjelaskan kondisi pasien dengan karno tipe 2.

```
# Asumsi Cox PH
cox.zph(model7)
##                chisq df       p
## karno_cluster  13.4  1 0.00026
## GLOBAL         13.4  1 0.00026
```

Dengan menguji asumsi Cox PH pada model 7 didapatkan seluruh variabel dan model 7 tidak memenuhi asumsi *proportional hazard* dengan *p-value* < 0,05.

**9) Model 8 (Menggunakan Semua Kovariat, Strata "celltype" dan "karno_cluster", *Backward* AIC)**

Uji Cox-PH dari model 8 yang menggunakan seluruh kovariat, strata pada variabel "celltype" dan "karno_cluster", serta *backward* AIC. Didapatkan tidak ada hasil yang signifikan oleh model 8.

## C. Perbandingan dan Informasi setiap Model

Berikut adalah tabel yang memberikan rangkuman dari tiap model yang telah dibuat

| Model | Penjelasan Model | Variabel Signifikan | AIC | Asumsi PH |
|---|---|---|---|---|
| 1 | Menggunakan semua kovariat | "celltypesmallcell", "celltypeadeno", "$karno_2$" | 979.8204 | Tidak terpenuhi |
| 2 | Menggunakan semua kovariat dan *backward* AIC | "celltypesmallcell", "celltypeadeno", "$karno_2$" | 974.8293 | Tidak terpenuhi |
| 3 | Menggunakan semua kovariat dan strata "karno_cluster" | "celltypesmallcell", "celltypeadeno" | 838.7498 | Tidak terpenuhi |
| 3.5 | Menggunakan semua kovariat, strata "karno_cluster", menghilangkan "prior". | "celltypesmallcell", "celltypeadeno" | 836.7514 | Terpenuhi |
| 4 | Menggunakan semua kovariat dan strata "celltype". | "$karno_2$" | 662.1234 | Tidak terpenuhi |

| | | | | |
|---|---|---|---|---|
| 5 | Menggunakan semua kovariat dan strata "celltype" serta "karno_cluster". | N/A | 532.1007 | Terpenuhi |
| 6 | Menggunakan semua kovariat, strata "karno_cluster", *backward* AIC. | "celltypesmallcell", "celltypeadeno" | 831.8335 | Terpenuhi |
| 7 | Menggunakan semua kovariat, strata "celltype", *backward* AIC. | "$karno_2$" | 657.0331 | Tidak terpenuhi |
| 8 | Menggunakan semua kovariat, strata "celltype" dan "karno_cluster", *backward* AIC. | N/A | N/A | Tidak terpenuhi |

Berdasarkan model-model yang memenuhi asumsi PH yaitu model 3, model 3.5, model 5, dan model 6. Dipilih model 6 untuk dilakukan uji lanjutan dengan mempertimbangkan nilai AIC yang rendah dan jumlah variabel yang signifikan. Jika diperhatikan model 5 memang memiliki AIC yang lebih kecil dibandingkan AIC pada model 6, namun model 6 tetap dipilih karena memiliki variabel yang signifikan yaitu "celltypesmall" dan "celltypeadeno".

Perhatikan lebih lanjut dari *summary* model 6 berikut

```
model6 <- stepAIC(object=model6_draft, data=veteran_data,
direction="backward")
summary(model6)

## Call:
## coxph(formula = Surv(time, status) ~ celltype + strata(karno_cluster),
##     data = veteran, method = "breslow")
##
##   n= 137, number of events= 128
##
##                     coef exp(coef) se(coef)     z Pr(>|z|)
## celltypesmallcell 0.8920    2.4400   0.2604 3.425 0.000615 ***
## celltypeadeno     1.0953    2.9902   0.2963 3.697 0.000218 ***
## celltypelarge     0.2833    1.3275   0.2796 1.013 0.310876
```

Dapat diinterpretasikan bahwa pasien dengan tipe sel squamous memiliki resiko kematian yang hampir sama dengan tipe sel adeno. Jika resiko kematian tertinggi diurutkan berdasarkan tipe sel maka didapatkan tipe sel adeno, squamous, adeno, dan large.

## D. Uji Lanjutan

Berdasarkan hasil uji asumsi Cox-PH pada model 6 berikut

```
# Asumsi Cox PH
cox.zph(model6)
##          chisq df    p
## celltype  7.42  3 0.06
## GLOBAL    7.42  3 0.06
```

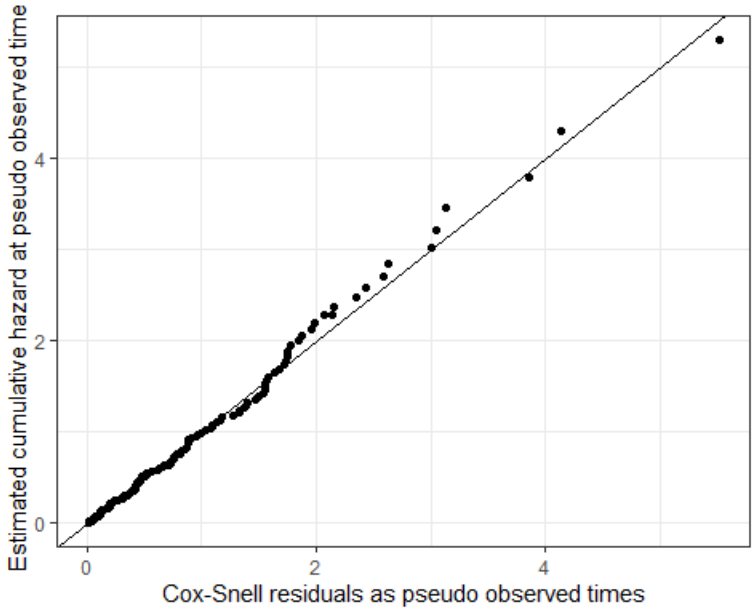
Dilakukanlah uji Grambsch dan Therneau sebagai pendekatan uji *goodness-of-fit* dengan nilai residual *Schoenfeld*. Uji ini dilakukan untuk mengevaluasi model memenuhi asumsi PH atau nilai β konstan sepanjang waktu. Dengan hipotesis null asumsi hazard terpenuhi dan *p-value* 0,06 > 0,05 maka didapatkan kesimpulan bahwa β konstan sepanjang waktu. Selanjutnya, model 6 akan diuji kembali dengan beberapa plot berikut.

### 1) Cox-Snell Residual Plot

Merupakan uji pada nilai residual kumulatif hazard yang ditunjukkan untuk mengetahui apakah model cocok secara keseluruhan tidak berfokus pada pelanggaran asumsi PH namun secara umum sesuai dengan data yang ada. Plot dari residual *cox-snell* ini akan mendekati garis lurus dengan *slope* 1 atau berdistribusi eksponensial dengan parameter 1. Secara matematik, residual *cox-snell* untuk individu ke-i dihitung dengan rumus

$$R_i = \widehat{H}_0(T_i)exp(\beta' X_i)$$

Berikut adalah hasil dari plot *cox-snell residual*



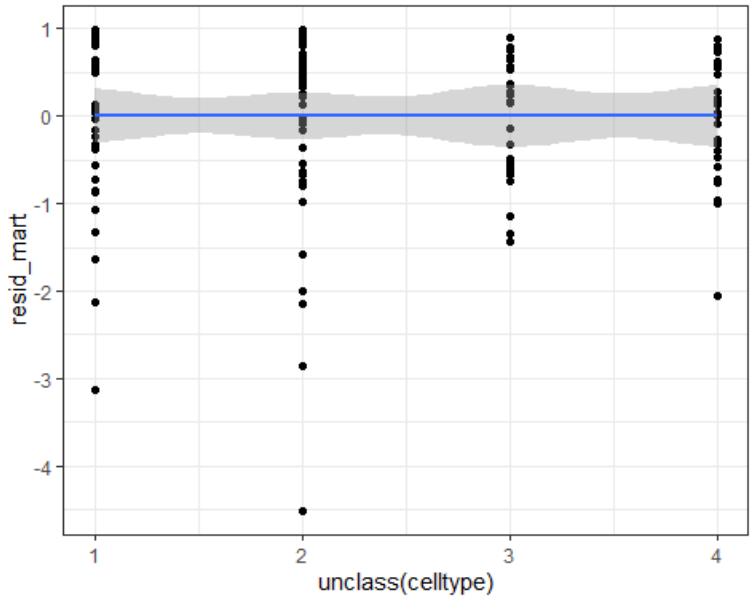

Terlihat bahwa titik-titik residual *cox-snell* berjalan mendekati garis 45 derajat yang mengindikasikan model memberikan kecocokan yang *reasonable*.

**2) Martingale Residual Plot**

Residual *martingale* adalah residual yang digunakan untuk mengevaluasi kecocokan pada model cox atau memeriksa apakah model cox yang diestimasi sesuai dengan data dan apakah kovariat yang tidak termasuk pada model harusnya ada. Plot ini digunakan untuk memeriksa asumsi linearitas dari kovariat. Nilai residual *martingale* yang negatif mengindikasikan data disensor dan residual *martingale* positif mengindikasikan terjadi kejadian. Secara matematis, residual *martingale* dinyatakan sebagai

$$M_i = \delta_i - \widehat{H}(t_i|X_i)$$

Berikut adalah plot dari residual *martingale*




Jika dilihat berdasarkan plot di atas, terlihat garis yang terbentuk lurus dan jatuh di sekitar nilai nol (tidak melebihi interval $\pm\, 0.5$) yang berarti terdapat kecenderungan asumsi linearitas pada kovariat terpenuhi.

# BAB IV PARTIAL LIKELIHOOD

Akan dibuat subset dari data berukuran n = 40 dan subset model dengan 1 prediktor kategorik dan 1 prediktor numerik. Pada kasus ini, digunakan variabel "trt" sebagai prediktor kategorik dan variabel "karno_cluster" sebagai prediktor numerik. Akan dipilih sembarang titik waktu terjadinya *event* dimana akan dilakukan konstruksi kontribusi *partial likelihood* dari titik waktu tersebut dan menuliskan bentuk fungsi *partial likelihood* secara keseluruhan.

## A. Data Tanpa *Ties*

Diambil subset data berukuran 40 secara acak dengan kondisi tanpa ties atau tidak terdapat event yang terjadi bersamaan. Berikut adalah beberapa baris pertama dari subset data tanpa *ties*.

```
> print(df_no_ties)
    trt karno time status
95    2    40    2      1
53    1    30    3      1
15    1    70   11      1
26    1    30   16      1
40    1    60   27      1
107   2    40   29      1
124   2    40   45      1
```

Dengan variabel kategorik "trt" sehingga dibentuklah variabel *dummy* sebagai berikut

$$T = \begin{cases} 1, test \\ . \\ 0, standard \end{cases}$$

Berdasarkan variabel *dummy* di atas didapatkan model $h(t) = h_0(t)\, exp(\beta_1 T + \beta_2\, karno)$

Sebelum mengonstruksi kontribusi *partial likelihood* dari satu titik waktu, terlebih dahulu akan dicari estimasi untuk $\beta_1$ dan $\beta_2$ dengan menggunakan metode *maksimum likelihood* dan didapatkan hasil berikut.

```
> cox_model <- coxph(Surv(time, status) ~ factor(trt) + karno, data = df_no_ties)
> # Melihat hasil model
> summary(cox_model)
Call:
coxph(formula = Surv(time, status) ~ factor(trt) + karno, data = df_no_ties)

  n= 40, number of events= 40

                coef exp(coef) se(coef)      z Pr(>|z|)
factor(trt)2 -0.37446   0.68766  0.38500 -0.973  0.33074
karno        -0.03212   0.96839  0.01020 -3.148  0.00164 **
---
```

Berdasarkan *summary* di atas, didapatkan nilai $\beta_1 = (-\ 0.37446)$ dan $\beta_2 = (-\ 0.03212)$ sehingga didapatkan model $h(t) = h_0(t)exp(-\ 0.37446treat\ -\ 0.03212\ karno_2)$

Konstruksi kontribusi *partial likelihood* pada satu titik waktu terjadinya *event* merupakan perbandingan relatif antara risiko yang dialami objek terhadap risiko dari seluruh objek yang ada pada waktu tersebut. Misalkan akan dikonstruksi partial *likelihood* pada waktu $t_1 = 2$ dan hanya terjadi 1 event dengan total risiko dari seluruh objek yang ada pada setiap waktu $t_j$ sebanyak 40 (19 *treatment standard* dan 21 *treatment test*). Perhatikan bahwa, pada $t_1 = 2$ risiko objek tersebut mengalami event adalah $h_0(2)\ e^{\beta_1 + \beta_2\ karno_2}$, sedangkan total risiko dari seluruh objek yang ada pada setiap waktu $t_j$ adalah $h_0(2)\ e^{\beta_1 + \beta_2\ karno_2} + h_0(3)\ e^{\beta_2\ karno_3} + \ \ldots\ + h_0(999)\ e^{\beta_1 + \beta_2\ karno_{999}}$

Sehingga, konstruksi partial likelihood pada saat $t_1 = 2$, yaitu

$$L(t_1) = \frac{h_0(2)\ e^{\beta_1 + \beta_2\ karno_2}}{h_0(2)\ e^{\beta_1 + \beta_2\ karno_2} + h_0(3)\ e^{\beta_2\ karno_3} + \ \ldots\ + h_0(999)\ e^{\beta_1 + \beta_2\ karno_{999}}}$$

$$= \frac{e^{\beta_1 + \beta_2\ karno_2}}{e^{\beta_1 + \beta_2\ karno_2} + e^{\beta_2\ karno_3} + \ \ldots\ + e^{\beta_1 + \beta_2\ karno_{999}}}$$

$$= \prod_{j=1}^{1} \frac{exp(\beta_1 T + \beta_2\ karno_2)}{\sum_{t_1}^{t_{40}} exp(\beta_1 T + \beta_2\ karno_{t_i})}$$

$$= \prod_{j=1}^{1} \frac{exp((-0.37446)T + (-0.03212)\ karno_2)}{\sum_{t_1}^{t_{40}} exp((-0.37446)T + (-0.03212)\ karno_{t_i})}$$

Sehingga didapatkan bentuk umum dari konstruksi *partial likelihood* untuk sembarang titik $t_j$, dengan D event dengan $t_1 < t_2 < \ \ldots \ < t_D$ adalah

$$L(t_j) = \prod_{j=1}^{D} \frac{exp(\beta_1 T + \beta_2 \ karno_2)}{\sum_{t_1}^{t_{40}} exp(\beta_1 T + \beta_2 \ karno_{t_i})}$$

$$= \prod_{j=1}^{D} \frac{exp((-0.37446)T + (-0.03212) \ karno_2)}{\sum_{t_1}^{t_{40}} exp((-0.37446)T + (-0.03212) \ karno_{t_i})}$$

Jika hanya ingin dipertimbangkan konstruksi *partial likelihood* berdasarkan salah satu variabel aja, maka didapatkan bentuk *likelihood* yang lebih sederhana terutama pada variabel kategorik. Misalkan akan dikonstruksi partial *likelihood* untuk variabel kategorik saja yaitu *treatment* pada waktu $t_1 = 2$ dan hanya terjadi 1 event dengan total risiko dari seluruh objek yang ada pada setiap waktu $t_j$ sebanyak 40 ( 19 *treatment standard* dan 21 *treatment test*). Perhatikan bahwa, pada $t_1 = 2$ risiko objek tersebut mengalami event adalah $h_0(2)\, e^{\beta_1}$, sedangkan total risiko dari seluruh objek yang ada pada setiap waktu $t_j$ adalah penjumlahan dari resiko 19 pasien untuk treatment standard dan 21 pasien dengan treatment test yaitu $19h_0(2) \ + 21h_0(2)\, e^{\beta_1}$.

Sehingga, konstruksi partial likelihood pada saat $t_1 = 2$, yaitu

$$L(t_1) = \frac{h_0(2)\, e^{\beta_1}}{19h_0(2) + 21h_0(2)\, e^{\beta_1}}$$

$$= \frac{e^{\beta_1}}{19 + 21\, e^{\beta_1}}$$

$$= \frac{e^{(-0.37446)}}{19 + 21\, e^{(-0.37446)}}$$

$$\approx 0.035$$

Sehingga didapatkan umum konstruksi *partial likelihood* untuk sembarang titik $t_j$, dengan D event dengan $t_1 < t_2 < \ldots < t_D$ adalah

$$L(t_j) = \prod_{j=1}^{D} \frac{exp(\beta_1 T)}{\sum_{t_1}^{t_{40}} exp(\beta_1 T)}$$

$$= \prod_{j=1}^{D} \frac{exp(\beta_1 T)}{19 + 21\, e^{\beta_1}}$$

Bentuk umum konstruksi *partial likelihood* di atas juga dapat digunakan dengan pendekatan yang sama untuk variabel numerik karno.

**B. Data dengan *Ties***

Diambil subset data berukuran 40 secara acak dengan kondisi tanpa *ties* atau tidak terdapat *event* yang terjadi bersamaan. Berikut adalah beberapa baris pertama dari subset data *ties*.

```
> print(df_ties)
    trt karno time status
77    2    20    1      1
85    2    50    1      1
95    2    40    2      1
53    1    30    3      1
18    1    40    4      1
42    1    50    7      1
97    2    20    7      1
119   2    40    7      1
12    1    40    8      1
46    1    20    8      1
```

Karena variabel "trt" merupakan variabel kategorik maka akan dibentuk variabel dummy berikut.

$$T = \begin{cases} 1, test \\ . \\ 0, standard \end{cases}$$

Sehingga model yang diajukan yaitu

$h(t) = h_0(t)\, exp(\beta_1 T + \beta_2\, karno)$

Sebelum mengonstruksi kontribusi partial likelihood dari satu titik waktu, terlebih dahulu akan dicari estimasi untuk $\beta_1$ dan $\beta_2$ dengan menggunakan metode maksimum likelihood.

```
> cox_model_ties <- coxph(Surv(time, status) ~ factor(trt) + karno, data = df_ties)
> # Melihat hasil model
> summary(cox_model_ties)
Call:
coxph(formula = Surv(time, status) ~ factor(trt) + karno, data = df_ties)

  n= 40, number of events= 39

                coef exp(coef) se(coef)      z Pr(>|z|)
factor(trt)2 -0.37765   0.68547  0.34695 -1.088    0.276
karno        -0.01662   0.98352  0.01027 -1.619    0.105
```

Sehingga didapat $\beta_1 = (-\ 0.37765)$ dan $\beta_2 = (-\ 0.01662)$ sehingga didapatkan model $h(t) = h_0(t)exp(-\ 0.37765treat\ -\ 0.01662\ karno_2)$.

Konstruksi kontribusi partial likelihood pada satu titik waktu terjadinya event merupakan perbandingan relatif antara risiko yang dialami objek terhadap risiko dari seluruh objek yang ada pada waktu tersebut. Misalkan akan dikonstruksi partial likelihood pada waktu $t = 8$ berdasarkan variabel kategorik saja yaitu treatment dengan menggunakan metode Breslow. Perhatikan bahwa pada waktu $t < 8$, terjadi 8 event dengan 3 event dialami pasien dengan treatment standard dan 5 event dialami pasien dengan treatment test. Sehingga pada saat $t = 8$ terdapat $40 - 8 = 32$ pasien pada himpunan risiko, yang terdiri dari $19 - 3 = 16$ pasien dengan *treatment standard* dan $21 - 5 = 16$ pasien dengan *treatment test*.

Total fungsi hazard untuk 16 pasien dengan *treatment standard* dan 16 pasien dengan *treatment test* adalah $16h_0(8) + 16h_0(8)e^{\beta_1}$. Dengan fungsi hazard untuk masing-masing pasien dengan treatment standard adalah $h_0(8)e^{\beta_1 .0} = h_0(8)$, sehingga kontribusi likelihood dari masing-masing pasien adalah

$$\frac{h_0(8)}{16h_0(8)+16h_0(8)e^{\beta_1}} = \frac{1}{16+16e^{\beta_1}}$$

Fungsi hazard untuk masing-masing pasien dengan *treatment test* adalah $h_0(8)e^{\beta_1 .1} = h_0(8)e^{\beta_1}$, sehingga kontribusi likelihood dari masing-masing pasien adalah $\frac{h_0(8)e^{\beta_1}}{16h_0(8)+16h_0(8)e^{\beta_1}} = \frac{e^{\beta_1}}{16+16e^{\beta_1}}$.

Diketahui terdapat 16 pasien dengan *treatment standard* dan 16 pasien dengan *treatment test* sehingga, kontribusi *likelihood* pada titik waktu $t = 8$ dengan Metode Breslow adalah

$$L(t = 8) = \frac{e^{\beta_1}}{(16+16e^{\beta_1})^{16}} x \frac{1}{(16+16e^{\beta_1})^{16}}$$

$$= \frac{e^{\beta_1}}{(16+16e^{\beta_1})^{32}}$$

$$= \frac{e^{(-0.37765)}}{(16+16e^{(-0.37765)})^{32}}$$

Didapatkan bentuk umum konstruksi *partial likelihood* untuk sembarang titik $t_j$ dengan menggunakan Metode Breslow yaitu

$$L(t = t_j) = \prod_{j=1}^{D} \frac{exp(\beta_1 T)}{\sum_{l \in R_i} exp(\beta_1 T)^{d_i}}$$

Dimana $l \in R_i$ menyatakan total pasien pada himpunan risiko untuk mengalami *event*, $d_i$ himpunan objek yang mengalami event di waktu $t_j$.

Pendekatan yang sama juga dapat digunakan jika ingin mengonstruksi partial likelihood baik untuk model lengkap maupun model dengan salah satu variabel (numerik atau kategorik saja).

# BAB V PENUTUP

## A. Kesimpulan

Pertama, dataset ‘veteran’ akan dicek *missing value* nya lalu akan dilakukan pengujian untuk variabel penjelas sehingga kita dapat memilih variabel penjelas yang signifikan. Dengan melihat visualisasi beberapa variabel kategorik, jika dilakukan pengujian log-rank terhadap fungsi survival untuk 2 jenis treatment, yakni $p = 0.93$ atau $p > 0.05$ sehingga variabel treatment bukanlah variabel yang signifikan untuk menjadi variabel penjelas, maka dilakukan pengujian log-rank terhadap fungsi survival untuk 4 jenis sel, yakni $p < 0.0001$ atau $p < 0.05$ sehingga variabel cell type merupakan salah satu variabel yang signifikan untuk menjadi variabel penjelas karena mengimplikasikan bahwa pasien dengan tipe sel squamous berisiko lebih besar untuk *survive*.

Kedua, akan di cek asumsi PH dengan grafis terlihat bahwa variabel trt, diagtime, age, dan prior memenuhi asumsi PH secara grafis sehingga dapat dilanjutkan ke pemodelan regresi *cox proportional hazard* yang nantinya bila bersumber dari buku model survival Ibu Sarini dilakukan pengajuan beberapa model lalu dilakukan pengujian asumsi Cox - PH di tiap model yang diajukan. Variabel lain seperti variabel *cell type* dan karno tidak memenuhi asumsi PH secara grafis yang artinya variabel bergantung pada waktu.

Ketiga, setelah diajukan beberapa model dan dilakukan pengujian asumsi Cox - PH di tiap model didapat model-model yang memenuhi asumsi PH yaitu model 3, model 3.5, model 5, dan model 6. Model 1, model 2, model 3, model 4, model 7, dan model 8 didapat hasil yang tidak memenuhi asumsi PH. Model 3 didapat hasil yang cukup baik namun masih terdapat variabel yang kurang signifikan seperti “celltype”. Model 3.5 didapat hasil yang cukup baik dengan semua variabel dan model 3.5 memenuhi asumsi PH. Model 5 didapat hasil yang cukup baik dengan semua variabel dan model 5 memenuhi asumsi PH. Model 6 didapat hasil yang cukup baik dengan semua variabel dan model 6 memenuhi asumsi PH. Model 6 terpilih untuk dilakukan uji lanjutan dengan mempertimbangkan nilai AIC yang rendah dan jumlah variabel yang signifikan. Jika diperhatikan model 5 memang memiliki AIC yang lebih kecil dibandingkan AIC pada model 6, namun model 6 tetap dipilih karena memiliki variabel yang signifikan yaitu

"celltypesmall" dan "celltypeadeno". Uji lanjutan model 6 dilakukan untuk mengevaluasi model memenuhi asumsi PH dan didapat hasil  yaitu model 6 memenuhi asumsi PH.

Terakhir, konstruksi kontribusi *partial likelihood*  40 data dengan kondisi tanpa ties pada saat $t_1 = 2$, yaitu

$$L(t_1) = \frac{h_0(2)\, e^{\beta_1}}{19h_0(2)+21h_0(2)\, e^{\beta_1}}$$

$$= \frac{e^{\beta_1}}{19 + 21\, e^{\beta_1}}$$

$$= \frac{e^{(-0.37446)}}{19 + 21\, e^{(-0.37446)}}$$

$$\approx 0.035$$

Bentuk umum konstruksi *partial likelihood* nya untuk sembarang titik $t_j$, dengan D event dengan $t_1 < t_2 < \ \ldots \ < t_D$ yaitu

$$L(t_j) = \prod_{j=1}^{D} \frac{exp(\beta_1 T)}{\sum_{t_1}^{t_{40}} exp(\beta_1 T)}$$

$$= \prod_{j=1}^{D} \frac{exp(\beta_1 T)}{19 + 21\, e^{\beta_1}}$$

Untuk variabel numerik saja yaitu karno juga dapat dilakukan pendekatan yang sama.

Selanjutnya, konstruksi kontribusi *partial likelihood*  40 data dengan kondisi terdapat ties pada saat $t = 8$ dengan Metode Breslow yaitu

$$L(t = 8) = \frac{e^{\beta_1}}{(16 + 16e^{\beta_1})^{16}}\ x\ \frac{1}{(16 + 16e^{\beta_1})^{16}}$$

$$= \frac{e^{\beta_1}}{(16 + 16e^{\beta_1})^{32}}$$

$$= \frac{e^{(-0.37765)}}{(16 + 16e^{(-0.37765)})^{32}}$$

Bentuk umum konstruksi partial likelihood nya untuk sembarang titik $t_j$ dengan menggunakan Metode Breslow yaitu

$$L(t = t_j) = \prod_{j=1}^{D} \frac{exp(\beta_1 T)}{\sum_{l \in R_i} exp(\beta_1 T)^{d_i}}$$

Dimana $l \in R_i$ menyatakan total pasien pada himpunan risiko untuk mengalami event, $d_i$ himpunan objek yang mengalami event di waktu $t_j$.

Pendekatan yang sama juga dapat digunakan jika ingin mengonstruksi partial likelihood baik untuk model lengkap maupun model dengan salah satu variabel (numerik atau kategorik saja).

## B. Saran

Saran yang dapat Tim penulis berikan dengan harapan dapat dikembangkan untuk penelitian selanjutnya adalah jika asumsi proporsional hazard dilanggar maka dapat digunakan sebagai alternatif yaitu Uji Straified Cox PHM.

# DAFTAR PUSTAKA


[1] Buana, Indra., Harahap, Dwi Agustian. (2022). ASBESTOS, RADON DAN POLUSI UDARA SEBAGAI FAKTOR RESIKO KANKER PARU PADA PEREMPUAN BUKAN PEROKOK- Jurnal Kedokteran dan Kesehatan Malikussaleh Vol.8 No.1 Mei 2022 https://ojs.unimal.ac.id/averrous/article/download/7088/3677

[2] Prabawati, Santi., Nasution, Yuki Novia., dan Wahyuningsih, Sri. (2018). Analisis Survival Data Kejadian Bersama dengan Pendekatan Efron Partial Likelihood (Studi Kasus: Lama Masa Studi Mahasiswa Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Mulawarman Angkatan 2011) - Jurnal EKSPONENSIAL Volume 9, Nomor 1, Mei 2018 https://jurnal.fmipa.unmul.ac.id/index.php/exponensial/article/download/278/130/

[3] D Kalbfleisch and RL Prentice (1980), The Statistical Analysis of Failure Time Data. Wiley, New York.

[4] Kleinbaum, D. G., & Klein, M. (2011). Survival Analysis: A Self-Learning Text, Third Edition (Statistics for Biology and Health). Springer.

[5] Schag CC, Heinrich RL, Ganz PA. Karnofsky performance status revisited: Reliability, validity, and guidelines. J Clin Oncology. 1984; 2:187-193.

[6] Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J., & Jemal, A. (2015). Global cancer statistics, 2012. CA: a cancer journal for clinicians, 65(2), 87–108. https://doi.org/10.3322/caac.21262

# LAMPIRAN

*Lampiran 1*

Codes Final Project Kelompok G Model Survival