Case Study: Cox Regression

Model Survival

Implementasi Cox Proportional Hazard dalam Pemodelan Survival Gagal Jantung

Aurelio Naufal Effendy (2106638526)

Rifqi Hafizuddin (2106638204)

Aina Grace Aritonang (2106651263)

Adawia Ananda (2106724883)

Michael Rich Kharisma (2106724901)

Myra Azzahra Putri Syah Indra (2106726844)

Nama	NPM	Kontribusi	Keaktifan
Aurelio Naufal Effendy	2106638526	Membantu mencari data, mengerjakan regresi Cox PH dan membantu menyusun kesimpulan.	100%
Rifqi Hafizuddin	2106638204	Mengerjakan pengecekan asumsi Proportional Hazard dan membantu menyusun kesimpulan.	100%
Aina Grace Aritonang	2106651263	Membantu penyusunan laporan dalam pembuatan kesimpulan.	100%
Adawia Ananda	2106724883	Mengerjakan latar belakang, menjelaskan dataset dan mendefinisikan variabel	100%
Michael Rich Kharisma	2106724901	Melakukan coding dan analisis untuk poin analisis deskriptif.	100%
Myra Azzahra Putri Syah Indra	2106726844	Mengerjakan latar belakang, menjelaskan dataset, dan mendefinisikan variabel.	100%

Daftar Isi

1. Latar Belakang

Penelitian ini meneliti gagal jantung untuk memprediksi risiko dan prognosis gagal jantung. Kumpulan data ini berisi informasi klinis dan demografis pasien yang didiagnosis gagal jantung, bersama dengan hasil kelangsungan hidup mereka. Penelitian ini untuk mengembangkan model prediktif dan mendapatkan wawasan tentang faktor-faktor yang mempengaruhi perkembangan gagal jantung dan kelangsungan hidup pasien.

Kumpulan data biasanya mencakup berbagai atribut yang merupakan prediktor potensial hasil gagal jantung. Atribut ini mungkin termasuk klinis, tubuh, dan faktor gaya hidup.

Para peneliti menggunakan kumpulan data ini untuk mengeksplorasi hubungan antara variabel-variabel ini dan perkembangan gagal jantung, serta untuk mengembangkan model yang dapat memperkirakan risiko terjadinya gagal jantung atau memprediksi waktu hingga kejadian terkait gagal jantung, seperti rawat inap atau kematian.

Untuk menganalisis dataset Heart Failure Prediction, digunakan model proporsional hazard Cox, yang memperkirakan rasio bahaya yang terkait dengan masing-masing prediktor, atau *machine learning* dapat memprediksi kelangsungan hidup pasien dengan gagal jantung hanya dari kreatinin serum dan fraksi ejeksi. Tujuannya adalah untuk mengidentifikasi prediktor yang signifikan dan membangun model akurat yang dapat membantu stratifikasi risiko dan manajemen pasien yang dipersonalisasi.

Dengan memanfaatkan kumpulan data Prediksi Gagal Jantung, maka akan meningkatkan pemahaman mereka tentang gagal jantung, mengidentifikasi individu berisiko tinggi yang mungkin mendapat manfaat dari intervensi dini, dan mengoptimalkan strategi pengobatan untuk meningkatkan hasil pasien.

2. Deskripsi Dataset

Dataset ini merupakan rekam medis dari 299 observasi gagal jantung dari pasien yang dikumpulkan dari Faisalabad Institute of Cardiology dan Allied Hospital Faisalabad (Punjab, Pakistan) pada interval waktu April-Desember 2015. Seluruh

pasien mengalami disfungsi sistolik ventrikel kiri dan pernah mengalami gagal jantung sebelumnya yang termasuk pada kelas III atau IV dari New York Heart Association.

Dataset ini terdiri dari 13 variabel yang mencakup informasi terkait klinis, tubuh, dan gaya hidup pasien. Variabel tersebut dijelaskan dalam tabel berikut:

Variabel	Jenis	Deskripsi
age	Numerik	Umur pasien
anaemia	Kategorik	Apakah pasien menderita anemia (Pengurangan sel darah merah) (0=tidak, 1=ya)
creatinine_phosphokinase	Numerik	Level dari enzim CPK pada darah (mcg/L)
diabetes	Kategorik	Apakah pasien menderita diabetes (0=tidak, 1=ya)
ejection_fraction	Numerik	Persentase darah meninggalkan jantung setiap kontraksi
high_blood_pressure	Kategorik	Apakah pasien menderita hipertensi (0=tidak, 1=ya)
platelets	Numerik	Jumlah platelet dalam darah (kiloplatelets/mL)
serum_creatinine	Numerik	Level kreatinin pada darah (mg/dL)
serum_sodium	Numerik	Level sodium pada darah (mEq/L)
sex	Kategorik	Jenis kelamin pasien (0=wanita, 1=pria)
smoking	Kategorik	Apakah pasien merokok (0=tidak, 1=ya)
time	Numerik	Periode tindak lanjut
death_event	Kategorik	Apakah pasien meninggal pada saat periode tindak lanjut (0=tidak, 1=ya)

3. Analisis Deskriptif

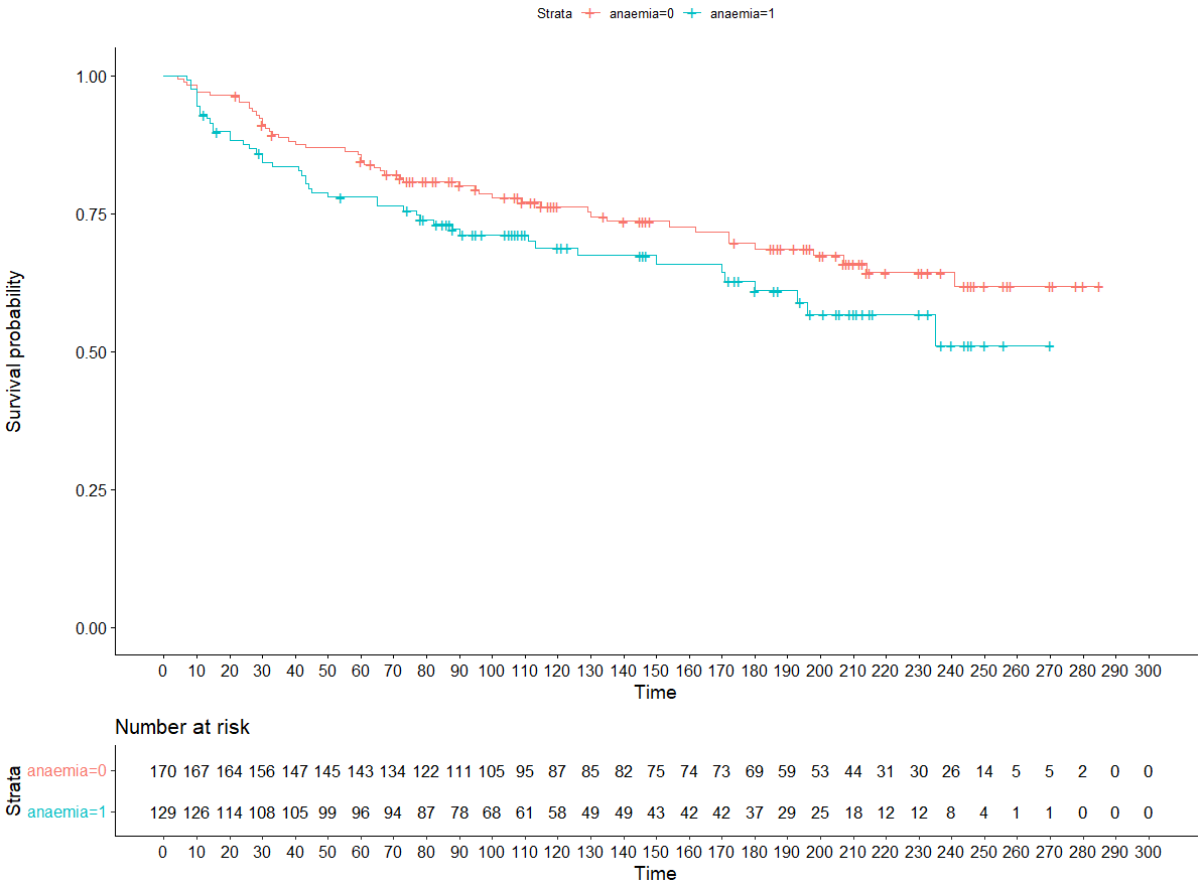
Akan dilakukan perbandingan survival berdasarkan variabel anaemia, diabetes, dan smoking.

Berikut code yang kami gunakan untuk variabel anaemia

```r
library(survival)
library(KMsurv)
library(survminer)
library(MASS)
km_model <- data %>%
  mutate(
    smoking = factor(ifelse(anaemia == 0, "non-anaemia", "anaemia"))
  ) %>%
  survfit(Surv(time, DEATH_EVENT) ~ anaemia, data = .)

ggsurvplot(km_model, data = data, risk.table = TRUE,
           break.time.by = 10, size = 0.3, tables.height = 0.20)
```

Berikut plot yang dihasilkan untuk variabel anaemia



Dari plot di atas maka dapat disimpulkan bahwa angka kematian dari penderita anaemia(biru) lebih tinggi atau kemungkinan survivenya lebih rendah dibanding yang tidak menderita anaemia(merah).

Berikut code yang kami gunakan untuk variabel diabetes

```r
library(survival)
library(KMsurv)
library(survminer)
library(MASS)
km_model <- data %>%
  mutate(
    smoking = factor(ifelse(diabetes == 0, "non-diabetes", "diabetes"))
  ) %>%
  survfit(Surv(time, DEATH_EVENT) ~ diabetes, data = .)

ggsurvplot(km_model, data = data, risk.table = TRUE,
           break.time.by = 10, size = 0.3, tables.height = 0.20)
```

Berikut plot yang dihasilkan untuk variabel diabetes

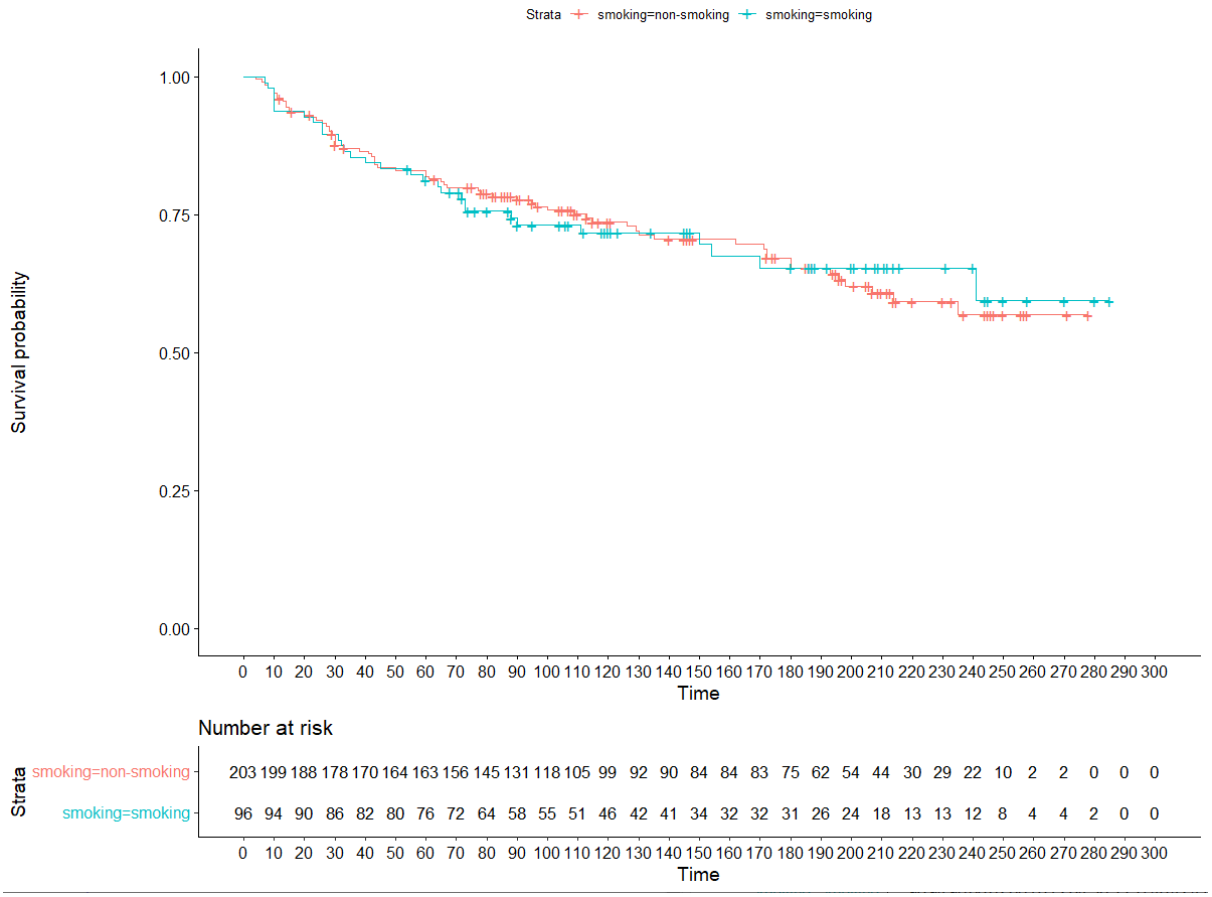
Dari plot di atas maka dapat disimpulkan bahwa diabetes tidak begitu memengaruhi angka kemungkinan survival tiap individu.

Berikut code yang kami gunakan untuk variabel smoking

```r
library(survival)
library(KMsurv)
library(survminer)
library(MASS)
km_model <- data %>%
  mutate(
    smoking = factor(ifelse(smoking == 0, "non-smoking", "smoking"))
  ) %>%
  survfit(Surv(time, DEATH_EVENT) ~ smoking, data = .)

ggsurvplot(km_model, data = data, risk.table = TRUE,
           break.time.by = 10, size = 0.3, tables.height = 0.20)
```

Berikut plot yang kami dapat untuk variabel smoking



Dari plot di atas maka dapat disimpulkan bahwa merokok tidak begitu memengaruhi angka kemungkinan survival tiap individu.

4. Regresi Cox - PH

Untuk memudahkan data untuk dimasukkan ke dalam program regresi Cox-PH, variabel-variabel kategorik perlu dilakukan pengkodean, atau pembentukan variabel dummy untuk kategori pada variabel-variabel kategorik tersebut dan menentukan baselinenya.

$$anaemic \begin{cases} 0, & nonanaemic \\ 1, & anaemic \end{cases}$$

$$diabetic \begin{cases} 0, & nondiabetic \\ 1, & diabetic \end{cases}$$

$$high\_blood\_pressure \begin{cases} 0, & non-high-bp \\ 1, & high-bp \end{cases}$$

$$sex \begin{cases} 0, & female \\ 1, & male \end{cases}$$

$$smoking \begin{cases} 0, & nonsmoker \\ 1, & smoker \end{cases}$$

Dengan pengkodean tersebut, yang menjadi baseline adalah yang bertanda 0 untuk masing-masing variabel kategorik, dengan program R :

```
#Ubah variabel yang sebelumnya boolean (0,1) menjadi variabel kategorik
#dan set levelnya (yang jadi base factor) untuk model fitting yang baik
df <- df %>%
  mutate(
    anaemia = factor(ifelse(anaemia == 1, "anaemic", "non-anaemic"), levels = c("non-anaemic", "anaemic")),
    diabetes = factor(ifelse(diabetes == 1, "diabetic", "non-diabetic"), levels = c("non-diabetic", "diabetic")),
    high_blood_pressure = factor(ifelse(high_blood_pressure == 1, "high-bp", "non-high-bp"), levels = c("non-high-bp", "high-bp")),
    sex = factor(ifelse(sex == 0, "female", "male"), levels = c("female", "male")),
    smoking = factor(ifelse(smoking == 0, "non-smoker", "smoker"), levels = c("non-smoker", "smoker")),
    platelets = platelets/1e4,
    creatinine_phosphokinase = creatinine_phosphokinase/1e3
  )
```

Sehingga model regresi Cox-PH yang diajukan adalah

$$\begin{aligned} h(t,x) = h_0(t)exp(&\beta_1 age + \beta_2 anaemia + \beta_3 creatinine\_phosphokinase + \beta_4 diabetes \\ &+ \beta_5 ejection\_fraction + \beta_6 high\_blood\_pressure + \beta_7 platelets \\ &+ \beta_8 serum\_creatinine + \beta_9 serum\_sodium + \beta_{10} smoking + \beta_{11} sex \end{aligned}$$

Dengan fungsi coxph pada program R, dengan n = 299 dan event yang terjadi 96, sisanya tersensor, diperoleh koefisiennya berdasarkan hasil R sebagai berikut :

$$\begin{aligned} h(t,x) = h_0(t)exp(&0.05age + 0.46anaemia + 0.22creatinine\_phosphokinase + 0.14diabetes \\ &- 0.04ejection\_fraction + 0.48high\_blood\_pressure - 0.005platelets \\ &+ 0.32serum\_creatinine - 0.04serum\_sodium + 0.13smoking - 0.24sex \end{aligned}$$

```
Call:
coxph(formula = Surv(time, DEATH_EVENT) ~ age + anaemia + creatinine_phosphokinas
e +
    diabetes + ejection_fraction + high_blood_pressure + platelets +
    serum_creatinine + serum_sodium + smoking + sex, data = df)

  n= 299, number of events= 96

                               coef exp(coef)  se(coef)      z Pr(>|z|)
age                        0.046408  1.047502  0.009324  4.977 6.45e-07
anaemiaanaemic             0.460135  1.584287  0.216837  2.122   0.0338
creatinine_phosphokinase   0.220737  1.246995  0.099185  2.225   0.0260
diabetesdiabetic           0.139884  1.150140  0.223147  0.627   0.5307
ejection_fraction         -0.048942  0.952237  0.010476 -4.672 2.98e-06
high_blood_pressurehigh-bp 0.475749  1.609219  0.216197  2.201   0.0278
platelets                 -0.004635  0.995376  0.011262 -0.412   0.6806
serum_creatinine           0.321032  1.378550  0.070170  4.575 4.76e-06
serum_sodium              -0.044188  0.956775  0.023266 -1.899   0.0575
smokingsmoker              0.128922  1.137602  0.251224  0.513   0.6078
sexmale                   -0.237521  0.788580  0.251609 -0.944   0.3452

age                        ***
anaemiaanaemic             *
creatinine_phosphokinase   *
diabetesdiabetic
ejection_fraction          ***
high_blood_pressurehigh-bp *
platelets
serum_creatinine           ***
serum_sodium               .
smokingsmoker
sexmale
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                           exp(coef) exp(-coef) lower .95 upper .95
age                           1.0475     0.9547    1.0285     1.067
anaemiaanaemic                1.5843     0.6312    1.0358     2.423
creatinine_phosphokinase      1.2470     0.8019    1.0267     1.515
diabetesdiabetic              1.1501     0.8695    0.7427     1.781
ejection_fraction             0.9522     1.0502    0.9329     0.972
high_blood_pressurehigh-bp    1.6092     0.6214    1.0534     2.458
platelets                     0.9954     1.0046    0.9736     1.018
serum_creatinine              1.3786     0.7254    1.2014     1.582
serum_sodium                  0.9568     1.0452    0.9141     1.001
smokingsmoker                 1.1376     0.8790    0.6953     1.861
sexmale                       0.7886     1.2681    0.4816     1.291

Concordance= 0.741  (se = 0.027 )
Likelihood ratio test= 81.95  on 11 df,   p=6e-13
Wald test            = 87.27  on 11 df,   p=6e-14
Score (logrank) test = 88.39  on 11 df,   p=3e-14
```

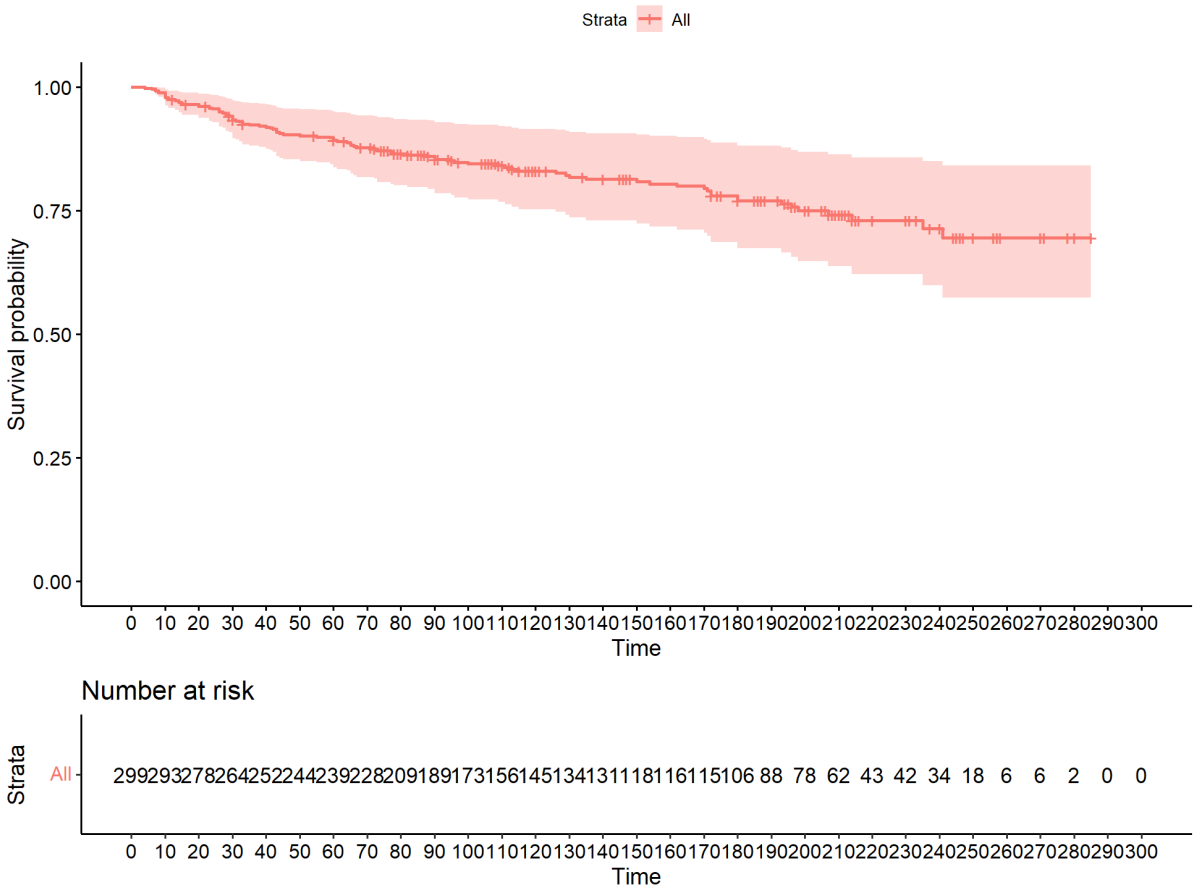
Terlihat bahwa pada pasien yang mengalami anaemia, hazard untuk mengalami infeksi adalah 1.5843 kali dibandingkan dengan hazard untuk pasien yang tidak mengalami anaemia. Hal ini mengimplikasikan bahwa pasien yang tidak mengalami anaemia memiliki resiko kematian yang lebih rendah daripada yang mengalami aanemia. Ini berlaku pula pada variabel-variabel kategorik lainnya. Lalu untuk

variabel numeriknya, berarti semakin tinggi nilai variabelnya, maka hazardnya lebih tinggi (tergantung pada nilai plus dan minus koefisiennya, jika minus berlaku kebalikan), contohnya dalam variabel ejection_fraction maka semakin tinggi nilai variabelnya, maka hazardnya makin rendah atau sebanyak 0.9522 kalinya.

Model ini cukup baik untuk menjelaskan beberapa penyebab dari pasien penyakit jantung meninggal. Terlihat dari uji model p-value untuk uji rasio likelihood adalah 6e-13, atau 0.00000000000006 yang mana sangat kecil dan jauh dibawah taraf signifikansi 0.05. Serta uji-uji lain juga memiliki p-valuenya yang signifikan. Hal ini mengimplikasikan bahwa variabel-variabel yang dimasukan dalam model tersebut dapat menjelaskan lama waktu hingga pasien meninggal. Berdasarkan uji parsial koefisien regresi (melihat p-valuenya), terlihat bahwa masing - masing pengukuran tersebut cukup signifikan dalam menjelaskan waktu kematian, kecuali untuk variabel diabetes, platelets, smoking dan sex.

Setelah model telah difit, kita dapat menggunakan model tersebut untuk memplot probabilitas survival kumulatif dari suatu populasi. Plot survival untuk populasi dengan nilai rata-rata kovariat regresi CoxPH adalah sebagai berikut :

```
ggsurvplot(survfit(cox_model), data = df, risk.table = TRUE, break.time.by = 10)
```



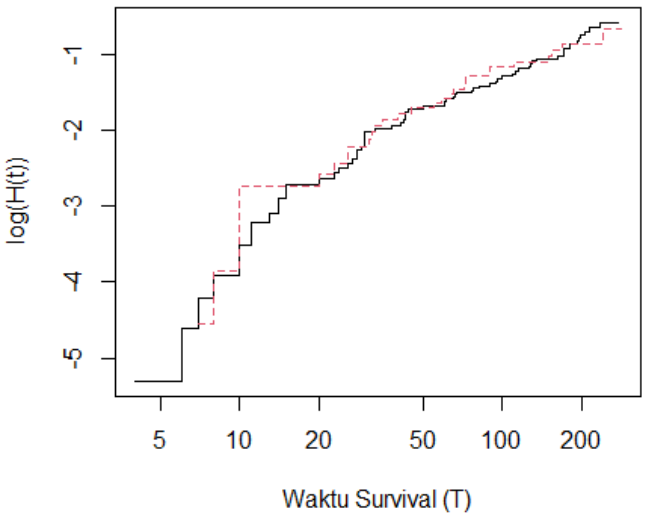
5. Pengecekan Asumsi Proportional Hazard

Akan dilakukan pengecekan asumsi PH pada variabel smoking, anemia, dan diabetes.

- Menggunakan Plot

a. smoking

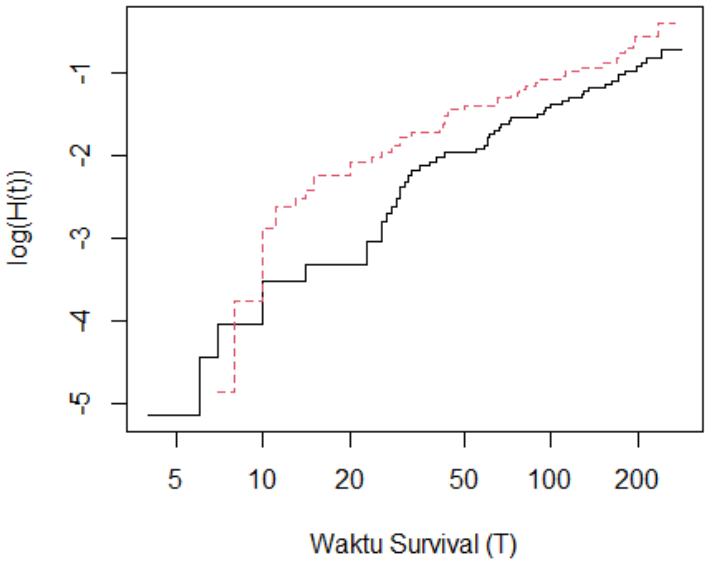
```
fit.df<- survfit(Surv(time, DEATH_EVENT) ~ smoking, data = df)
plot(fit.df,fun="cloglog",lty=1:2,col=1:2, mark.time=FALSE,xlab="Waktu Survival (T)", ylab="log(H(t))")
```



Dengan menggunakan kode tersebut, diperoleh plot perbandingan tipe “smoking” dan “non-smoking”. Terlihat bahwa fungsi *log(H(t))* pada variabel smoking relatif paralel, yang mengindikasikan terpenuhinya asumsi PH untuk variabel smoking.

b. anemia

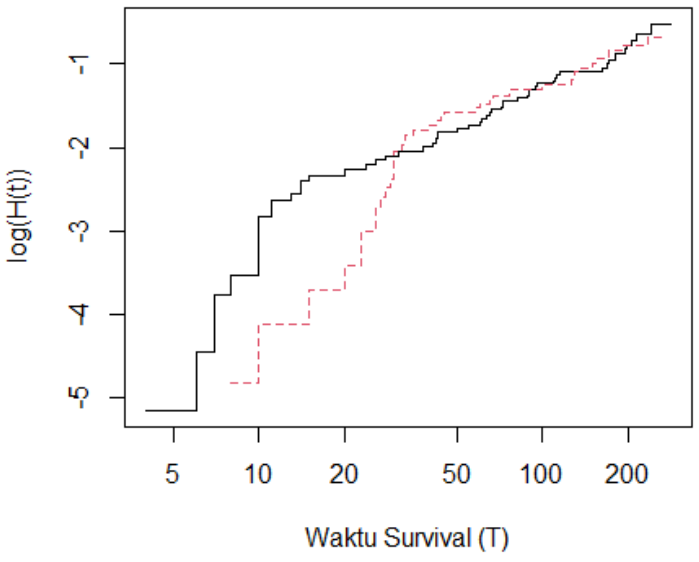
```
fit2.df<- survfit(Surv(time, DEATH_EVENT) ~ anaemia, data = df)
plot(fit2.df,fun="cloglog",lty=1:2,col=1:2, mark.time=FALSE,xlab="Waktu Survival (T)", ylab="log(H(t))")
```



Dengan menggunakan kode tersebut, diperoleh plot perbandingan tipe “anemic” dan “non-anemic”. Terlihat bahwa fungsi *log(H(t))* pada variabel anemia relatif paralel, yang mengindikasikan terpenuhinya asumsi PH untuk variabel anemia.

c. diabetes

```
fit3.df<- survfit(Surv(time, DEATH_EVENT) ~ diabetes, data = df)
plot(fit3.df,fun="cloglog",lty=1:2,col=1:2, mark.time=FALSE,xlab="Waktu Survival (T)", ylab="log(H(t))")
```



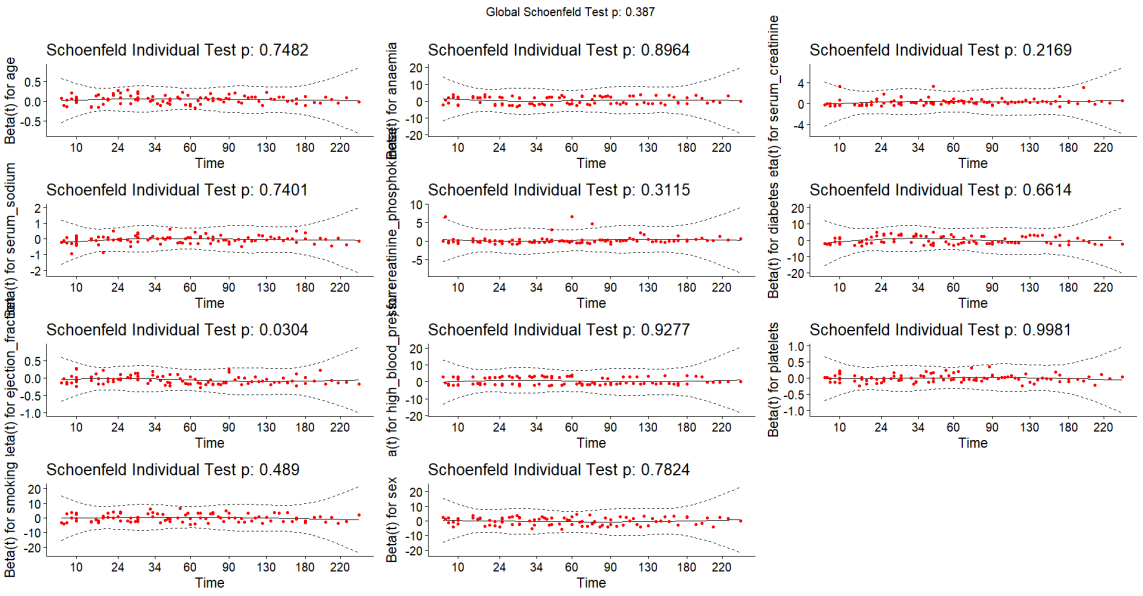
Dengan menggunakan kode tersebut, diperoleh plot perbandingan tipe “diabetic” dan “non-diabetic”. Terlihat bahwa fungsi *log(H(t))* pada variabel diabetes relatif paralel, yang mengindikasikan terpenuhinya asumsi PH untuk variabel diabetes.

- Menggunakan *scaled schoenfeld residual test*

Asumsi besar dari model proportional hazard adalah bahwa setiap kovariat hanya dapat mendongkrak atau menekan baseline hazard secara proporsional. Artinya, koefisien untuk setiap kovariat tidak bervariasi terhadap waktu. Hal tersebut dapat diuji menggunakan *scaled schoenfeld residual test.*

```
> cox.zph(cox_model)
                            chisq df    p
age                      1.03e-01  1 0.75
anaemia                  1.69e-02  1 0.90
serum_creatinine         1.52e+00  1 0.22
serum_sodium             1.10e-01  1 0.74
creatinine_phosphokinase 1.02e+00  1 0.31
diabetes                 1.92e-01  1 0.66
ejection_fraction        4.69e+00  1 0.03
high_blood_pressure      8.23e-03  1 0.93
platelets                5.69e-06  1 1.00
smoking                  4.79e-01  1 0.49
sex                      7.63e-02  1 0.78
GLOBAL                   1.17e+01 11 0.39
```

Tes tersebut melaporkan statistik uji chi-square untuk semua kovariat + untuk model global. Kami mengamati bahwa Global tidak signifikan pada tingkat 5% dan Hanya ejection_fraction yang signifikan pada level 5%. Kita harus memplot residu Schoenfeld untuk membuat keputusan:



Berdasarkan plot, terlihat bahwa tidak ada kovariat yang memiliki residu yang bervariasi waktu. Asumsi PH berlaku.

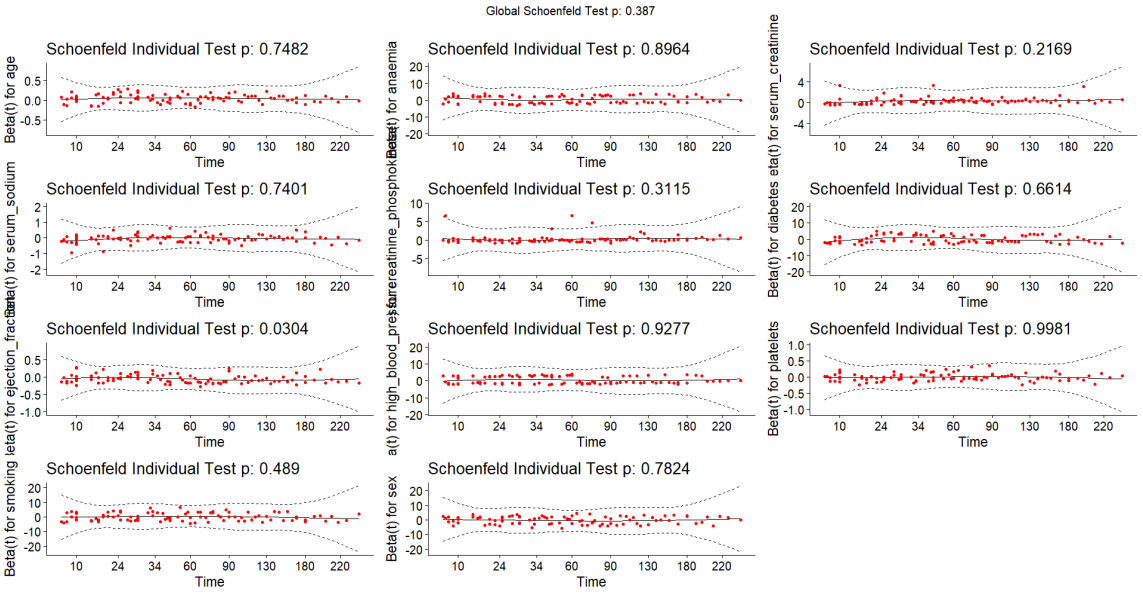
6. Kesimpulan

Setelah melakukan pemodelan Cox Proportional hazard terhadap kasus gagal jantung, didapat fungsi hazard proporsional sebagai berikut,

$$h(t,x) = h_0(t)exp(0.05age + 0.46anaemia + 0.22creatinine\_phosphokinase + 0.14diabetes - 0.04ejection\_fraction + 0.48high\_blood\_pressure - 0.005platelets + 0.32serum\_creatinine - 0.04serum\_sodium + 0.13smoking - 0.24sex$$

Model ini cukup baik dalam menjelaskan beberapa penyebab dari pasien penyakit gagal jantung meninggal, variabel-variabel yang dimasukan dalam model tersebut dapat menjelaskan lama waktu hingga pasien meninggal. Berdasarkan uji parsial koefisien regresi, didapatkan bahwa masing-masing pengukuran tersebut cukup signifikan dalam menjelaskan waktu kematian, kecuali untuk variabel diabetes, platelets, smoking dan sex.

Serta dari plot Schoenfeld untuk membuat keputusan didapat:



Berdasarkan plot, terlihat bahwa tidak ada kovariat yang memiliki residu Schoenfeld yang bervariasi waktu sehingga asumsi PH/proportional hazard berlaku.

7. Referensi

Chicco, D., & Jurman, G. (2020, February 3). *Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone - BMC Medical Informatics and decision making*. BioMed Central. https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5

8. Lampiran

a. Syntax R untuk Analisis Deskriptif

```
library(dplyr)
library(psych)
library(survminer)

# Read the data
data <- read.csv("heart_failure_clinical_records_dataset.csv")

head(data)

# Summary statistics
summary(data)

# Descriptive statistics for each variable
describe(data)

# Frequency table for a categorical variable
table(data$categorical_variable)

# Histogram for a numeric variable
hist(data$numeric_variable)

# Box plot for a numeric variable
boxplot(data$numeric_variable)

# Correlation matrix
cor(data)

# Scatter plot for two numeric variables
plot(data$numeric_variable1, data$numeric_variable2)

ggsurvplot(St.Z1, conf.int=TRUE, pval=TRUE, risk.table=TRUE,
legend.labs=c("0=age","1=anaemia","2=creatinine_phosphokinase","3=diabetes","4=
ejection_fraction","5=high_blood_pressure","6=platelets","7=serum_creatinine","8=s
erum_sodium
","9=sex","10=smoking","11=time","12=DEATH_EVENT")legend.title="Z1",palette
=c"dodgerblue1","orchid1"),title="", risk.table.height=.3)

library(survival)
library(KMsurv)
library(survminer)
library(MASS)
km_model <- data %>%
  mutate(
    smoking = factor(ifelse(smoking == 0, "non-smoking", "smoking"))
  ) %>%
```

```
  survfit(Surv(time, DEATH_EVENT) ~ smoking, data = .)

ggsurvplot(km_model, data = data, risk.table = TRUE,
           break.time.by = 10, size = 0.3, tables.height = 0.20)


library(survival)
library(KMsurv)
library(survminer)
library(MASS)
km_model <- data %>%
  mutate(
    smoking = factor(ifelse(diabetes == 0, "non-diabetes", "diabetes"))
  ) %>%
  survfit(Surv(time, DEATH_EVENT) ~ diabetes, data = .)

ggsurvplot(km_model, data = data, risk.table = TRUE,
           break.time.by = 10, size = 0.3, tables.height = 0.20)


library(survival)
library(KMsurv)
library(survminer)
library(MASS)
km_model <- data %>%
  mutate(
    smoking = factor(ifelse(anaemia == 0, "non-anaemia", "anaemia"))
  ) %>%
  survfit(Surv(time, DEATH_EVENT) ~ anaemia, data = .)

ggsurvplot(km_model, data = data, risk.table = TRUE,
           break.time.by = 10, size = 0.3, tables.height = 0.20)
```

b. Syntax R untuk Regresi Cox - PH

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

#import dataset
df <- heart_failure_clinical_records_dataset

#Ubah variabel yang sebelumnya boolean (0,1) menjadi variabel kategorik
#dan set levelnya (yang jadi base factor) untuk model fitting yang baik
df <- df %>%
  mutate(
```

```
  anaemia = factor(ifelse(anaemia == 1, "anaemic", "non-anaemic"), levels =
c("non-anaemic", "anaemic")),
  diabetes = factor(ifelse(diabetes == 1, "diabetic", "non-diabetic"), levels =
c("non-diabetic", "diabetic")),
  high_blood_pressure = factor(ifelse(high_blood_pressure == 1, "high-bp",
"non-high-bp"), levels = c("non-high-bp", "high-bp")),
  sex = factor(ifelse(sex == 0, "female", "male"), levels = c("female",
"male")),
  smoking = factor(ifelse(smoking == 0, "non-smoker", "smoker"), levels =
c("non-smoker", "smoker")),
  platelets = platelets/1e4,
  creatinine_phosphokinase = creatinine_phosphokinase/1e3
 )

#cek data
df %>% head

# Cox proportional hazard model
cox_model <- coxph(Surv(time, DEATH_EVENT) ~ age + anaemia +
creatinine_phosphokinase + diabetes + ejection_fraction +
                    high_blood_pressure + platelets + serum_creatinine +
serum_sodium + smoking + sex,
                  data = df)
summary(cox_model)

# Karena model telah fit, kita dapat menggunakannya untuk memplot
probabilitas survival kumulatif dari suatu populasi.
ggsurvplot(survfit(cox_model), data = df, risk.table = TRUE, break.time.by =
10)
```

c. Syntax R untuk Pengecekan Asumsi Proportional Hazard

```
df<-read.csv("C:\\Users\\rifqi\\Documents\\Semester 4\\Model
Survival\\heart_failure_clinical_records_dataset.csv")



library(survival)
library(KMsurv)
library(survminer)
library(MASS)

#import dataset
```

```
#Ubah variabel yang sebelumnya boolean (0,1) menjadi variabel kategorik
#dan set levelnya (yang jadi base factor) untuk model fitting yang baik
df <- df %>%
  mutate(
    anaemia = factor(ifelse(anaemia == 1, "anaemic", "non-anaemic"), levels =
c("non-anaemic", "anaemic")),
    diabetes = factor(ifelse(diabetes == 1, "diabetic", "non-diabetic"), levels =
c("non-diabetic", "diabetic")),
    high_blood_pressure = factor(ifelse(high_blood_pressure == 1, "high-bp",
"non-high-bp"), levels = c("non-high-bp", "high-bp")),
    sex = factor(ifelse(sex == 0, "female", "male"), levels = c("female",
"male")),
    smoking = factor(ifelse(smoking == 0, "non-smoker", "smoker"), levels =
c("non-smoker", "smoker")),
    platelets = platelets/1e4,
    creatinine_phosphokinase = creatinine_phosphokinase/1e3
  )

#cek data
df %>% head

# Cox proportional hazard model
cox_model <- coxph(Surv(time, DEATH_EVENT) ~ age + anaemia +
serum_creatinine + serum_sodium +creatinine_phosphokinase + diabetes +
ejection_fraction +
                     high_blood_pressure + platelets + smoking + sex,
                   data = df)
summary(cox_model)

# Plot the survival for a population with mean value of covariates
ggsurvplot(survfit(cox_model), data = df, risk.table = TRUE, break.time.by =
10)


fit.df<- survfit(Surv(time, DEATH_EVENT) ~ smoking, data = df)
plot(fit.df,fun="cloglog",lty=1:2,col=1:2, mark.time=FALSE,xlab="Waktu
Survival (T)", ylab="log(H(t))",)

fit2.df<- survfit(Surv(time, DEATH_EVENT) ~ anaemia, data = df)
plot(fit2.df,fun="cloglog",lty=1:2,col=1:2, mark.time=FALSE,xlab="Waktu
Survival (T)", ylab="log(H(t))")

fit3.df<- survfit(Surv(time, DEATH_EVENT) ~ diabetes, data = df)
```

```
plot(fit3.df,fun="cloglog",lty=1:2,col=1:2, mark.time=FALSE,xlab="Waktu
Survival (T)", ylab="log(H(t))")
```

```
options(repr.plot.width = 18, repr.plot.height = 12)
ggcoxzph(cox.zph(cox_model))
cox.zph(cox_model)
```