UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

BIOSTATISTICS II

ASSIGNMENT 3

2022

STUDENT NAME: Nokwanda Themba

STUDENT NO.: THMNOK003

COURSE CODE: PPH7092S

DUE DATE: 3 October 2022

Total [50 Marks]

THIS ASSIGNMENT INCLUDES::

1. Software used: R
2. Data Set: BiostatII_2022_Assign3_data.51.csv
3. THMNOK003_R Script Assignment 3

LIST OF APPENDICES:

I. R Script Code File

Question 1 [20 marks]: Non-parametric approach (Descriptive statistics)

a) **Plot Kaplan-Meier curves** for: [8]

i) the **different number of comorbidities strata**

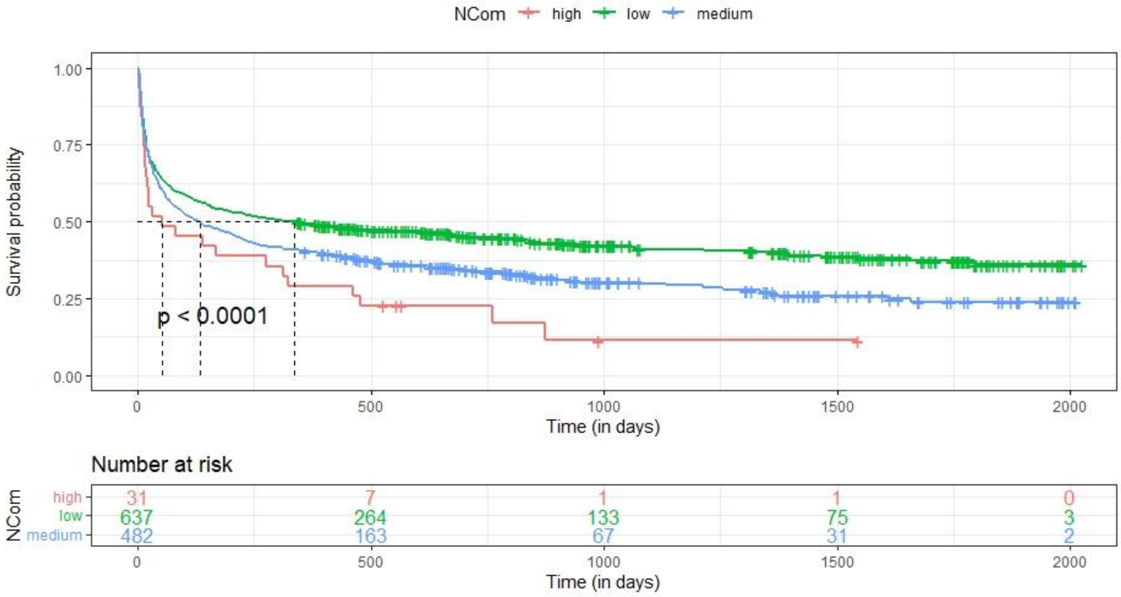

Figure 1. Kaplan-Meier curves displaying the estimated survival probabilities for the different number of comorbidities of adult individuals with critical, acute conditions such as acute respiratory failure and multiple organ system failure

ii) different **monthly income strata (Figure 2)**

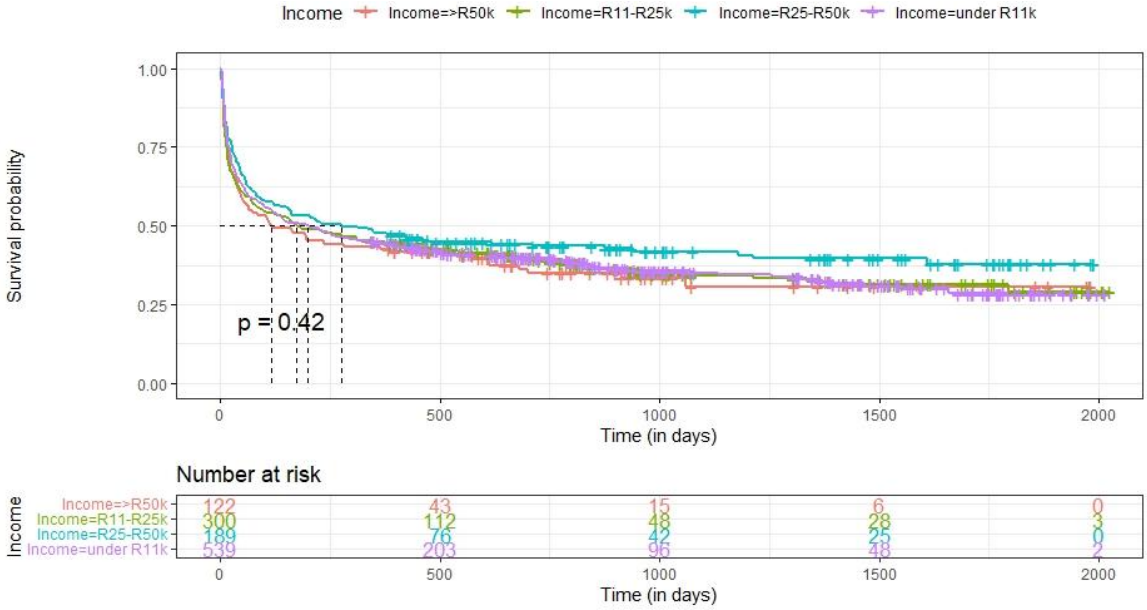

Figure 2. Kaplan-Meier curves displaying the estimated survival probabilities for the different monthly income strata of adult individuals with critical, acute conditions such as acute respiratory failure and multiple organ system failure

b) **Compare** the **survival curves** of the groups below **using the log-rank test**. State the hypothesis being tested and **briefly interpret** the results. [8]

Null hypothesis:
$\mathbf{H_0}$**:** There is no difference in terms of survival or the distribution of time until the event between the two groups (i.e. S1t = S2t).

Alternative hypothesis:
$\mathbf{H_A}$**:** There is a difference in terms of survival or the distribution of time until the event between the two groups/survival in the two groups is not the same (i.e. S1t ≠ S2t)

i) the **number of comorbidities:**

Test statistics: Chi-Squared test statistic is χ2 = 18.8 on 2 degrees of freedom and the corresponding p-value is <0.001 (p-value < 0.05) which is less than some significance level (i.e α = .05)

We therefore have enough statistical evidence to reject the null hypothesis. We have enough statistical evidence to reject the null and conclude that there is sufficient evidence to say there is a statistically significant difference in survival between the two groups.

Through visual inspection (Figure 1.) we can see that the lines are not touching each other. The groups have different distribution curves. We conclude that the survival probabilities for adult individuals with low number of comorbidities are significantly higher survival than the survival probabilities for adult individuals with medium and high number of comorbidities

ii) the **monthly income group indicator**

Test statistics: Chi-Squared test statistic is χ2= 2.8 on 3 degrees of freedom and the corresponding p-value is = 0.42 (p-value > 0.05) which is greater than some significance level (i.e. α = .05)

We therefore do not have sufficient statistical evidence to reject the null hypothesis. We fail to reject the null hypothesis, we don't have sufficient evidence to conclude that there is a statistically significant difference in survival between the two groups (Both groups have similar distribution curves.)

Through visual inspection (Figure 2.) we can see that the lines are touching each other. The groups have similiar distribution curves. We conclude that the survival probabilities for adult individuals with a monthly income between R25-R50k are higher than the survival probabilities for adult individuals with a monthly income under R11k, R11-R25k and over R50k

c) Based on the plots in **Q1a (i) and (ii)**: **provide estimates of the survival probabilities** at **365 days** with **95% confidence intervals for each group**. **Interpret the results** (ONE SHORT paragraph) [4]

(i) the **number of comorbidities:**

NCom	Survival Probability	95% CI
high	0.29	(0.17, 0.50)
low	0.49	(0.46, 0.53)
medium	0.40	(0.36, 0.45)

(ii) the **monthly income group indicator**:

Income	Survival Probability	95% CI
>R50k	0.43	(0.35, 0.52)
R11-R25k	0.45	(0.40, 0.51)
R25-R50k	0.48	(0.41 0.55)
under R11k	0.45	(0.41, 0.49)

Interpretation:

The probability of an adult with a low number of comorbidities surviving for a year (i.e. 365 days) is 0.29; for an adult with a medium number of comorbidities, the probability of surviving for 365 days is 0.40; and for an adult with a low number of comorbidities, the probability of surviving for 365 days is 0.29. Compared to those with low and medium numbers of comorbidities, those with a high number of comorbidities had the lowest chances of survival. Adults with lower comorbidities had the highest probability of survival. The probability of survival decreases the higher number of comorbidities an adult individual has.

The probability of an adult with a monthly income of over >R50k surviving for a year (i.e. 365 days) is 0.43; for an adult with a monthly income of under R11k, the probability of surviving for 365 days is 0.45; for an adult with a monthly income of R11-R25k, the probability of surviving for 365 days is 0.45; and and for an adult with a monthly income of R25-R50k, the probability of surviving for 365 days is 0.48. Compared to those with a monthly income of R11-R25k, R25-R50k, and under R11k, those with a monthly income of over >R50k had the lowest chances of survival. Adults with a monthly income of R25-R50k had the highest probability of survival. The probability of survival increase seems to be highest at an adult individual's monthly income of R25-R50k.

Question 2 [30 marks]: **Semi-parametric approach (Modelling)**

a) Present a single table **(Table 1)** showing the results of the **models** below with **Hazard Ratios (HR)**, **95% confidence intervals and p-values**. [15]

i) **Univariable Cox model**, for the **covariates age, number of comorbidities, monthly income, mean arterial pressure, heart rate and serum creatine**.

ii) **Multivariable Cox model** investigating the **effect** of **number of comorbidities on the primary outcome, adjusting for age, monthly income, mean arterial pressure, heart rate and serum creatine as potential confounders.**

Table 1: Univariate and Multivariate Cox proportional hazard models with covariates age, number of comorbidities, monthly income, mean arterial pressure, heart rate and serum creatine and time until death as outcome (n= 1150)

Variables	**Univariate analysis**		**Multivariate analysis**	
	HR (95% CI)	**p-value**	**HR (95% CI)**	**p-value**
Age (years)	1.02 (1.01 - 1.02)	<0.001	1.02 (1.01-1.02)	<0.001
NCom				
high	-	-	-	-
low	0.54 (0.36 - 0.80)	<0.001	0.56 (0.37-0.83)	<0.001
medium	0.70 (0.47 - 1.05)	0.08	0.71 (0.47 - 1.06)	0.09
Income (rands)				
under R11k	0.93 (0.73 - 1.19)	0.6	0.89 (0.69 - 1.14)	0.4
R11-R25k	0.95 (0.73 - 1.24)	0.7	0.97 (0.74 - 1.26)	0.8
R25-R50k	0.80 (0.60 - 1.07)	0.14	0.81 (0.61 - 1.09)	0.2
>R50k	-	-	-	-
MAP (mmHg)	1.00 (0.99 - 1.00)	0.01	1.00 (1.00 - 1.00)	0.09
Pulse (beats/minute)	1.00 (1.00 - 1.01)	<0.001	1.00 (1.00 - 1.01)	<0.001
Crea (mg/dL)	1.07 (1.04 - 1.10)	<0.001	1.06 (1.02 - 1.09)	<0.001

1 HR-hazard ratio; CI-confidence interval; NCom-Number of comorbidities; MAP-Mean arterial pressure (mmHg); Pulse-Heart rate (beats/minute); Crea-Serum creatinine (mg/dL)

b) Completely **interpret your Cox regression analysis.** In conclusion, how did the effect of the number of comorbidities change after adjusting for other covariates? [15]

The primary outcome was time until death of adult individuals with critical, acute conditions in a multi-hospital study. We assessed the risk of experiencing the event at a time ‘t’ for covariates age, number of comorbidities, monthly income, mean arterial pressure, heart rate and serum creatine.

In our **univariable cox model**, for the covariates age, number of comorbidities, monthly income, mean arterial pressure, pulse and serum creatine.

Global statistical significance of each model: In our univariable cox models we can see that the p-value for all three overall tests (likelihood, Wald, and score) are significant (all p-values were p=<0.001, respectfully [*p*<.05]) for predictors age, serum, pulse and number of comorbidities, indicating that the model is significant.

For predictors Income and MAP the overall tests (likelihood, Wald, and score) were not significant (the p-values were (p=0.4 and p=0.01 respectively). [*p*>.05]

For the covariate age, the p-value is *p*<0.001 (p <.05), with a hazard ratio of *HR (95% CI)* =1.02 (1.01 - 1.02) indicating a significant relationship between the age of adult individuals at the beginning of the study and our primary outcome, time until death. A 1-unit increase in age increases the hazard of death by 1.02 times. This is significantly different from 0 (p-value<.0.05) (i.e. HR > 1=increased hazard)

For the covariate serum creatinine the p-value is *p*<0.001 (p <.05), with a hazard ratio of *HR (95% CI)* =1.07 (1.04 - 1.10)) indicating a significant relationship between the serum creatinine levels of adult individuals and our primary outcome, time until death. A 1-unit increase in serum creatinine increases the hazard of death by 1.07 times. This is significantly different from 0 (p-value<.0.05)(i.e. HR > 1 = increased hazard).

For the covariate pulse the p-value is *p*=0.004 (p <.05), with a hazard ratio of *HR (95% CI)* =1.00 (1.00 - 1.01) indicating a significant relationship between the pulse of adult individuals and our primary outcome, time until death. A 1-unit increase in pulse has no effect on survival. This is significantly different from 0 (p-value<.0.05) (i.e.HR = 1: no effect)

For the covariate MAP the p-value for is *p* =0.009 (p <.05), with a hazard ratio of *HR (95% CI)* =1.00 (0.99 - 1.00) indicating a significant relationship between the MAP of adult individuals and our primary outcome, time until death. A 1-unit increase in MAP has no effect on survival. This is significantly different from 0 (p-value<.0.05) (i.e.HR = 1: no effect)

For the covariate low number of comorbidities, the p-value is *p*<0.001 (*p* <.05), with a hazard ratio of *HR (95% CI)* =0.54 (0.36 - 0.80) indicating a significant relationship between the adult individuals with a low number of comorbidities and our primary outcome, time until death. For adult individuals with a low number of comorbidities, the hazard of death would be 46% lower than those with high or medium number of comorbidities. This is significantly different from 0 (p-value<.0.05) (i.e. HR < 1 = reduction in hazard)

For the covariate monthly income of under R11k the p-value for is *p*=0.6 (*p* >.05), with a hazard ratio of *HR (95% CI)* =0.93 (0.73 - 1.19) indicating no significant relationship between adult individuals with a monthly income of under R11k and our primary outcome, time until death. For adults with a monthly income of under R11k, the hazard of death is 7% lower than those earning R11-R25k, R25-R50K and >R50k per month. This is not significantly different from 0 (p-value>.0.05) (i.e. HR < 1 = reduction in hazard)

For the covariate monthly income of R11-R25k the p-value for is *p*=0.7 (*p* >.05), with a hazard ratio of *HR (95% CI)* = 0.95 (0.73 - 1.24) indicating no significant relationship between adult individuals with a monthly income of R11-R25k and our primary outcome, time until death. For adults with a monthly income of R11-R25k, the hazard of death is 5% lower than those earning under R11k, R25-R50K and >R50k per month. This is not significantly different from 0 (p-value>.0.05) (i.e. HR < 1 = reduction in hazard)

For the covariate monthly income of R25-R50k the p-value for is *p*=0.14 (*p* >.05), with a hazard ratio of *HR (95% CI)* =0.80 (0.60 - 1.07) indicating no significant relationship between adult individuals with a monthly income of R25-R50 and our primary outcome, time until death. For adults with a monthly income of R25-R50, the hazard of death is 20% lower than those earning under R11k, R11-R25k and >R50k per month. This is not significantly different from 0 (p-value>.0.05) (i.e. HR < 1 = reduction in hazard)

For the covariate meidum number of comorbidities, the p-value is *p*=0.08 (*p* >.05), with a hazard ratio of *HR (95% CI)* =0.70 (0.47 - 1.05) indicating no significant relationship between the adult individuals with a medium number of comorbidities and our primary outcome, time until death. For adult individuals with a medium number of comorbidities, the hazard of death would be 30% lower than those with high or low number of comorbidities. This is not significantly different from 0 (p-value>.0.05) (i.e. HR < 1 = reduction in hazard)

In our **multivariable cox model** we were interested in investigating the effect of number of comorbidities on the primary outcome, adjusting for age, monthly income, mean arterial pressure, pulse and serum creatine as potential confounders

Global statistical significance of the model: Looking at our multivariable cox model we can see that the p-value for all three overall tests (likelihood, Wald, and score) are significant (p=<0.001), indicating that the model is significant.

The corresponding p-values of the coefficients for the following variables: all monthly income strata (i.e under R11k,R11-R25k,R25-R50k) were *p*=0.4, *p*= 0.8 *and p*=0.2, respectfully, for MAP (*p*=0.09) and for medium number of comorbidities (*p*=0.09) were not significant (*p-value* >.05) while the corresponding p-values for the coefficients for age were (*p*<0.001), heart rate (*p* <0.001), serum creatinine (*p*<0.001) and low number of comorbidities (*p*<0.001) which were statistically significant (*p-value* <.05).

The p-value for age is *p*<0.001 (p <.05), with a hazard ratio of *HR (95% CI)* =1.02 *(1.01-1.02)* indicating a significant relationship between the age of adult individuals at the beginning of the study and our primary outcome, time until death. Holding other covariates constant, a higher value of age is associated with poor survival (i.e. HR > 1=increased hazard, predictor age is associated with increased risk and decreased survival)

The p-value for pulse is *p* <0.001 (p <.05), with a hazard ratio of *HR (95% CI)* =1.00 (1.00 - 1.01) indicating a significant relationship between the pulse of adult individuals and our primary outcome, time until death. Holding other covariates constant, a higher pulse has no effect on survival (i.e.HR = 1: no effect, predictor pulse does not affect survival)

The p-value for MAP is *p* = 0.093 (p >.05), with a hazard ratio of *HR (95% CI)* =1.00 (1.00 - 1.00) indicating no significant relationship between the MAP of adult individuals and our primary outcome, time until death. Holding other covariates constant, a higher value of MAP has no effect on survival (i.e.HR = 1: no effect, predictor MAP does not affect survival)

The p-value for serum creatinine is *p* <0.001 (p <.05), with a hazard ratio of *HR (95% CI)* =1.06 *(1.02 - 1.09)* indicating a significant relationship between the serum creatinine levels of adult individuals and our primary outcome, time until death. Holding other covariates constant, a higher value of serum creatinine is associated with poor survival (i.e. HR > 1 = increased hazard, predictor serum creatinine is associated with increased risk and decreased survival).

The p-value for low number of comorbidities is *p* <0.001 (*p* <.05), with a hazard ratio of *HR (95% CI)* =0.56 (0.37-0.83) indicating a significant relationship between the adult individuals with a low number of comorbidities and our primary outcome, time until death. Holding other covariates constant, a higher value of serum creatinine is associated with improved survival (i.e. HR < 1 = reduction in hazard, predictor low number of comorbidities is associated with decreased risk and improved survival)

The p-value for medium number of comorbidities is *p*=0.09 (*p* >.05), with a hazard ratio of *HR (95% CI)* =0.71 *(0.47 - 1.06)* indicating no significant relationship between the adult individuals with a medium number of comorbidities and our primary outcome, time until death. Holding other covariates constant, a medium number of comorbidities is associated with improved survival (i.e. HR < 1 = reduction in hazard, predictor medium number of comorbidities is associated with decreased risk and improved survival)

The p-value for a monthly income of under R11k is *p*=0.4 (*p* >.05), with a hazard ratio of *HR (95% CI)* =0.89 (0.69 - 1.14) indicating no significant relationship between adult individuals with a monthly income of under R11k and our primary outcome, time until death. Holding other covariates constant, a monthly income of under R11k is associated with improved survival (i.e. HR < 1 = reduction in hazard, predictor monthly income of under R11k is associated with decreased risk and improved survival)

The p-value for a monthly income of R11-R25k is *p*=0.8 (*p* >.05), with a hazard ratio of *HR (95% CI)* =0.97 (0.74 - 1.26) indicating no significant relationship between adult individuals with a monthly income of R11-R25k and our primary outcome, time until death. Holding other covariates constant, a monthly income of R11-R25k is associated with improved survival (i.e. HR < 1 = reduction in hazard, predictor monthly income of R11-R25k is associated with decreased risk and improved survival)

The p-value for a monthly income of R25-R50k is *p*=0.2 (*p* >.05), with a hazard ratio of *HR (95% CI)* =0.81 (0.61 - 1.09) indicating no significant relationship between adult individuals with a monthly income of R25-R50k and our primary outcome, time until death. Holding other covariates constant, a monthly income of R25-R50k is associated with improved survival (i.e. HR < 1 = reduction in hazard, predictor monthly income of R25-R50k is associated with decreased risk and improved survival).

Conclusion:

In both our univariable and multivariable cox model, in regards to the individual p-values, covariates age, pulse, serum creatinine and low number of comorbidities were all statistically significant (p <.05) and 'useful' in predicting the primary outcome, time until death.

Adjusting for age, monthly income, mean arterial pressure, heart rate and serum creatine as potential confounders, a low number of comorbidities has a significant effect on the primary outcome, time until death. Predictors low and medium number of comorbidities are associated with decreased risk/hazard and improved survival.

APPENDIX 1

```
install.packages("gtsummary")
install.packages("tidyverse")
install.packages("funModeling")
install.packages("lubridate")
install.packages("KMsurv")
install.packages("survival")
install.packages("survminer")
install.packages("broom")
install.packages("here")
install.packages("plotly")
install.packages("epitools")
install.packages("kfigr")
install.packages("pacman")
library("gtsummary")
library("tidyverse")
library("funModeling")
library("lubridate")
library("KMsurv")
library("survival")
library("survminer")
library("broom")
library("here")
library("plotly")
library("epitools")
library("kfigr")
library("pacman")
##exploring data
View(BiostatII_2022_Assign3_data.51)
glimpse(BiostatII_2022_Assign3_data.51)
head(BiostatII_2022_Assign3_data.51)
tail(BiostatII_2022_Assign3_data.51)
summary(BiostatII_2022_Assign3_data.51)
row.names(BiostatII_2022_Assign3_data.51)
str(BiostatII_2022_Assign3_data.51)
##table summary
BiostatII_2022_Assign3_data.51 %>%
  tbl_summary(by = Death)
# distribution of survival times
ggplot(BiostatII_2022_Assign3_data.51, aes(x = Time, color = as.factor(Death))) +
  geom_histogram(
    fill = "white",
    stat = "bin",
    binwidth = 1) +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(title = "Status of the individuals") +
  xlab("Time to death (years)") +
  ylab("Frequency")  +
  guides(fill = guide_legend(title = "Death"),
         colour = guide_legend(title = "Death"))
##Kaplan-Meier plot
plot(survival::survfit(survival::Surv(Time, Death == 1) ~ 1, data =
BiostatII_2022_Assign3_data.51),
     xlab = "Time (Days)",
     ylab = "Survival probability")
##
BiostatII_2022_Assign3_data.51 %>% survminer::ggsurvplot(
  xlab = "Time (Days)",
  ylab = "Survival Probability"#,
  #font.title = c(11, "dark red"),
  #censor = TRUE,
  #conf.int = TRUE,
  #risk.table = TRUE,
  #risk.table.y.text = TRUE,
  #risk.table.height = 0.25,
  #surv.median.line = "hv"
)
#KM Model
km.model <- survfit(Surv(Time, Death) ~ 1, type = "kaplan-meier",
                     data = BiostatII_2022_Assign3_data.51)
km.model_2 <- survfit(Surv(Time, Death) ~ NCom, type = "kaplan-meier",
                    data = BiostatII_2022_Assign3_data.51)
km.model_3 <- survfit(Surv(Time, Death) ~ Income, type = "kaplan-meier",
                      data = BiostatII_2022_Assign3_data.51)
##summary of model
km.model
km.model_2
km.model_3
summary(km.model)
summary(km.model_2)
summary(km.model_3)
tidy(km.model) %>% head()
tidy(km.model_2) %>% head()
tidy(km.model_3) %>% head()
plot(km.model)
plot(km.model_2)
plot(km.model_3)
##adding confidence intervals, red line,  around the survival function
plot(km.model, conf.int = T, xlab = "Time (days)", ylab = "Alive = S(t)", main =
"Kaplan Meier Model 1", las = 1, mark.time = TRUE)
abline (h=0.5, col= "red")
ggsurvplot(km.model)

plot(km.model_2, conf.int = T, xlab = "Time (days)", ylab = "Alive = S(t)", main =
"Kaplan Meier Model 1", las = 1, mark.time = TRUE)
abline (h=0.5, col= "red")
ggsurvplot(km.model_2)

plot(km.model_3, conf.int = T, xlab = "Time (days)", ylab = "Alive = S(t)", main =
"Kaplan Meier Model 1", las = 1, mark.time = TRUE)
abline (h=0.5, col= "red")
ggsurvplot(km.model_3)
##Estimating the Survival Function - S(t)
##Creating the survival object
survival::Surv(BiostatII_2022_Assign3_data.51$Time,
BiostatII_2022_Assign3_data.51$Death)[1:10]
##The life table method for estimating S(t)
cuts <- seq(0, 21, 3)
BiostatII_2022_Assign3_data.51 %>%
  mutate(Time = cut(Time, cuts)) %>%
  group_by(Time) %>%
  summarise(ncensor = sum(Death == 0),
            nevent = sum(Death == 1))-> lifetab_df
##Kaplan-Meier estimator for S(t)
kp_plot_com <- survival::survfit(survival::Surv(Time, Death == 1) ~ 1, data =
BiostatII_2022_Assign3_data.51)
names(kp_plot_com)
```

```
print(kp_plot_com)
tidy(kp_plot_com) %>% head()
##Kaplan-Meier plot - base R
plot(survival::survfit(survival::Surv(Time, Death == 1) ~ 1, data =
BiostatII_2022_Assign3_data.51),
     xlab = "Time (Days)",
     ylab = "Survival probability")
##Kaplan-Meier plot- ggsurvplot
kp_plot_com %>% survminer::ggsurvplot(
  xlab = "Time (days)",
  ylab = "Survival Probability")
##Calculating survival time and censoring
Surv(time = BiostatII_2022_Assign3_data.51$Time, event =
BiostatII_2022_Assign3_data.51$Death)
view(BiostatII_2022_Assign3_data.51)
##analysis of event of interest
survfit(Surv(time = Time, event = Death) ~ 1, data =
BiostatII_2022_Assign3_data.51)
survfit(Surv(time = Time, event = Death) ~ NCom, data =
BiostatII_2022_Assign3_data.51)
survfit(Surv(time = Time, event = Death) ~ Income, data =
BiostatII_2022_Assign3_data.51)
##
s1 = survfit(Surv(time = Time, event = Death) ~ 1, data =
BiostatII_2022_Assign3_data.51)
s2 = survfit(Surv(time = Time, event = Death) ~ NCom, data =
BiostatII_2022_Assign3_data.51)
s3 = survfit(Surv(time = Time, event = Death) ~ Income, data =
BiostatII_2022_Assign3_data.51)
summary(s1)
summary(s2)
summary(s3)
##
plot(s1)
plot(s2)
plot(s3)

ggsurvplot(s1)
ggsurvplot(s2)
ggsurvplot(s3)
##
##Adding risk table
ggsurvplot(s1, xlab = "Time (in days)",
           ylab = "Survival probability",
           surv.median.line = "hv",
           risk.table = TRUE)

ggsurvplot(s2, xlab = "Time (in days)",
           ylab = "Survival probability",
           surv.median.line = "hv",
           risk.table = TRUE)

ggsurvplot(s3, xlab = "Time (in days)",
           ylab = "Survival probability",
           surv.median.line = "hv",
           risk.table = TRUE)

##changing variable
##NCom
```

```
s2_ = survfit(Surv(time = Time, event = Death) ~ NCom, data =
BiostatII_2022_Assign3_data.51)
s2_
ggsurvplot(s2_, xlab = "Time (in days)",
           ylab = "Survival probability",
           surv.median.line = "hv",
           risk.table = TRUE)
#Income
s3_ = survfit(Surv(time = Time, event = Death) ~ Income, data =
BiostatII_2022_Assign3_data.51)
s3_
ggsurvplot(s3_, xlab = "Time (in days)",
           ylab = "Survival probability",
           surv.median.line = "hv",
           risk.table = TRUE)
##Adding confidence interval
ggsurvplot(s1_, xlab = "Time (in days)",
           ylab = "Survival probability",
           surv.median.line = "hv",conf.int = TRUE,
           risk.table = TRUE, pval = TRUE)
ggsurvplot(s2_, xlab = "Time (in days)",
           ylab = "Survival probability",
           surv.median.line = "hv",legend.title = "NCom",conf.int = TRUE,
           risk.table = TRUE, pval = TRUE)
##Customized survival curves
ggsurvplot(s2_, xlab = "Time (in days)",
           ylab = "Survival probability",
           surv.median.line = "hv",legend.title = "NCom",conf.int = TRUE,
           risk.table = TRUE, pval = TRUE, size = 1,        # Add risk table
           risk.table.col = "strata",legend.labs =
             c("high", "low","medium"),risk.table.height = 0.25, ggtheme =
theme_bw())
##Without CI
ggsurvplot(s2_, xlab = "Time (in days)",
           ylab = "Survival probability",
           surv.median.line = "hv",legend.title = "NCom",
           risk.table = TRUE, pval = TRUE, size = 1,        # Add risk table
           risk.table.col = "strata",legend.labs =
             c("high", "low","medium"),risk.table.height = 0.25, ggtheme =
theme_bw())
##
ggsurvplot(s3_, xlab = "Time (in days)",
           ylab = "Survival probability",
           surv.median.line = "hv",legend.title = "Income",conf.int = TRUE,
           risk.table = TRUE, pval = TRUE)
##Customized survival curves
ggsurvplot(s3_, xlab = "Time (in days)",
           ylab = "Survival probability",
           surv.median.line = "hv",legend.title = "Income",conf.int = TRUE,
           risk.table = TRUE, pval = TRUE, size = 1,        # Add risk table
           risk.table.col = "strata",legend.labs =
             c(">R50k", "R11-R25k","R25-R50k","under R11k"),risk.table.height =
0.25, ggtheme = theme_bw())
##Without CI
ggsurvplot(s3_, xlab = "Time (in days)",
           ylab = "Survival probability",
           surv.median.line = "hv",legend.title = "Income",
           risk.table = TRUE, pval = TRUE, size = 1,        # Add risk table
           risk.table.col = "strata",risk.table.height = 0.25, ggtheme =
theme_bw())
## GG Plot Customized survival curves
ggsurvplot(
  s2_,
  data = BiostatII_2022_Assign3_data.51,
  size = 1,                 # change line size
  palette =
    c("#E7B800", "#2E9FDF"),# custom color palettes
  conf.int = TRUE,          # Add confidence interval
  pval = TRUE,              # Add p-value
  risk.table = TRUE,        # Add risk table
  risk.table.col = "strata",# Risk table color by groups
  legend.labs =
    c("high", "low","medium"),    # Change legend labels
  risk.table.height = 0.25, # Useful to change when you have multiple groups
  ggtheme = theme_bw()      # Change ggplot2 theme
)
##Log Rank Tests
survdiff(Surv(time = Time, event = Death) ~ NCom,data =
BiostatII_2022_Assign3_data.51)
survdiff(Surv(time = Time, event = Death) ~ Income,data =
BiostatII_2022_Assign3_data.51)
##we can choose to plot simple curves
plot(survfit(Surv(time = Time, event = Death) ~ NCom, data =
BiostatII_2022_Assign3_data.51),
     xlab = "Time",
     ylab = "Overall survival probability")
plot(survfit(Surv(time = Time, event = Death) ~ Income, data =
BiostatII_2022_Assign3_data.51),
     xlab = "Time",
     ylab = "Overall survival probability")
#### The life table method for estimating S(t)
cuts <- seq(0, 21, 3)
BiostatII_2022_Assign3_data.51 %>%
  mutate(time_cat = cut(Time, cuts)) %>%
  group_by(time_cat) %>%
  summarise(ncensor = sum(Death == 0),
            nevent = sum(Death == 1))-> lifetab_df
with(lifetab_df,
     lifetab(tis = cuts, ninit = nrow(BiostatII_2022_Assign3_data.51), nlost =
ncensor, nevent = nevent)) -> df_lifetab
round(df_lifetab, 3)
#### The Kaplan-Meier estimator for S(t)
##For NCom
KM_s2 <- survival::survfit(survival::Surv(Time, Death == 1) ~ ~ NCom, data =
BiostatII_2022_Assign3_data.51)
names(KM_s2)
print(KM_s2)
tidy(KM_s2) %>% head()
##For Income
KM_s3 <- survival::survfit(survival::Surv(Time, Death == 1) ~ Income, data =
BiostatII_2022_Assign3_data.51)
names(KM_s3)
print(KM_s3)
tidy(KM_s3) %>% head()
##estimates of the survival probabilities
##For NCom
print(s2_)
summary(s2_, times = 365)
```

```
surv_s2 <-survfit(Surv(time = Time, event = Death) ~ NCom, data =
BiostatII_2022_Assign3_data.51)
summary(surv_s2,times = 365)
tidy(s2_) %>% tail()
#For Income
print(s3_)
summary(s3_, times = 365)
surv_s3 <-survfit(Surv(time = Time, event = Death) ~ Income, data =
BiostatII_2022_Assign3_data.51)
summary(surv_s3,times = 365)
tidy(s3_) %>% tail()
##cox proportional hazard
##Multivariate analysis
Cox_multi <- coxph(Surv(Time, Death == 1) ~  Age + NCom + Income + MAP + Pulse +
Crea, data = BiostatII_2022_Assign3_data.51)
summary(Cox_multi)
##graphically looking at cox results
ggforest(Cox_multi)
##Tabulating
Cox_multi %>%
  tbl_regression(
    exp = TRUE)
mUltivariate_tab <-tbl_regression(Cox_multi, exponentiate = TRUE,conf.int = TRUE)
mUltivariate_tab
#Univariate Cox Models
##Age
Cox_age <- coxph(Surv(Time, Death == 1) ~  Age, data =
BiostatII_2022_Assign3_data.51)
summary(Cox_age)
tidy(Cox_age)
broom::tidy(Cox_age,exponentiate= TRUE,conf.int = TRUE)
Cox_agetab <- tbl_regression(Cox_age, exponentiate = TRUE,conf.int = TRUE)
Cox_agetab
##NCom
Cox_NCom <- coxph(Surv(Time, Death == 1) ~  NCom, data =
BiostatII_2022_Assign3_data.51)
summary(Cox_NCom)
tidy(Cox_NCom)
broom::tidy(Cox_NCom,exponentiate= TRUE,conf.int = TRUE)
Cox_NComtab <- tbl_regression(Cox_NCom, exponentiate = TRUE,conf.int = TRUE)
Cox_NComtab
##Income
Cox_Income <- coxph(Surv(Time, Death == 1) ~  Income, data =
BiostatII_2022_Assign3_data.51)
summary(Cox_Income)
tidy(Cox_Income)
broom::tidy(Cox_Income,exponentiate= TRUE,conf.int = TRUE)
Cox_Incometab <- tbl_regression(Cox_Income, exponentiate = TRUE,conf.int = TRUE)
Cox_Incometab
##MAP
Cox_MAP<- coxph(Surv(Time, Death == 1) ~  MAP, data =
BiostatII_2022_Assign3_data.51)
summary(Cox_MAP)
tidy(Cox_MAP)
broom::tidy(Cox_MAP,exponentiate= TRUE,conf.int = TRUE)
Cox_MAPtab <- tbl_regression(Cox_MAP, exponentiate = TRUE,conf.int = TRUE)
Cox_MAPtab
##Pulse
Cox_Pulse <- coxph(Surv(Time, Death == 1) ~  Pulse, data =
BiostatII_2022_Assign3_data.51)
summary(Cox_Pulse)
tidy(Cox_Pulse)
broom::tidy(Cox_Pulse,exponentiate= TRUE,conf.int = TRUE)
Cox_Pulsetab <- tbl_regression(Cox_Pulse, exponentiate = TRUE,conf.int = TRUE)
Cox_Pulsetab
##Crea
Cox_Crea <- coxph(Surv(Time, Death == 1) ~  Crea, data =
BiostatII_2022_Assign3_data.51)
summary(Cox_Crea)
tidy(Cox_Crea)
broom::tidy(Cox_Crea,exponentiate= TRUE,conf.int = TRUE)
Cox_Creatab <- tbl_regression(Cox_Crea, exponentiate = TRUE,conf.int = TRUE)
Cox_Creatab
##Combining both tables
##Combine Univariate results
tbl_merge(
  tbls =
list(Cox_agetab,Cox_NComtab,Cox_Incometab,Cox_MAPtab,Cox_Pulsetab,Cox_Creatab),)
Univariate_tab <- tbl_merge(
  tbls =
list(Cox_agetab,Cox_NComtab,Cox_Incometab,Cox_MAPtab,Cox_Pulsetab,Cox_Creatab),)
Univariate_tab
##combine univariate and mutlivariate models in a single table
tbl_merge(
  tbls = list(Univariate_tab, mUltivariate_tab),
  tab_spanner = c("**Univariate_Model**", "**Multivariate_Model**"))
##Test the proportional hazard assumption
plot(cox.zph(Cox_multi))
cox.zph(Cox_multi)
ggcoxdiagnostics(Cox_multi,'schoenfeld',ox.scale = 'time')
```