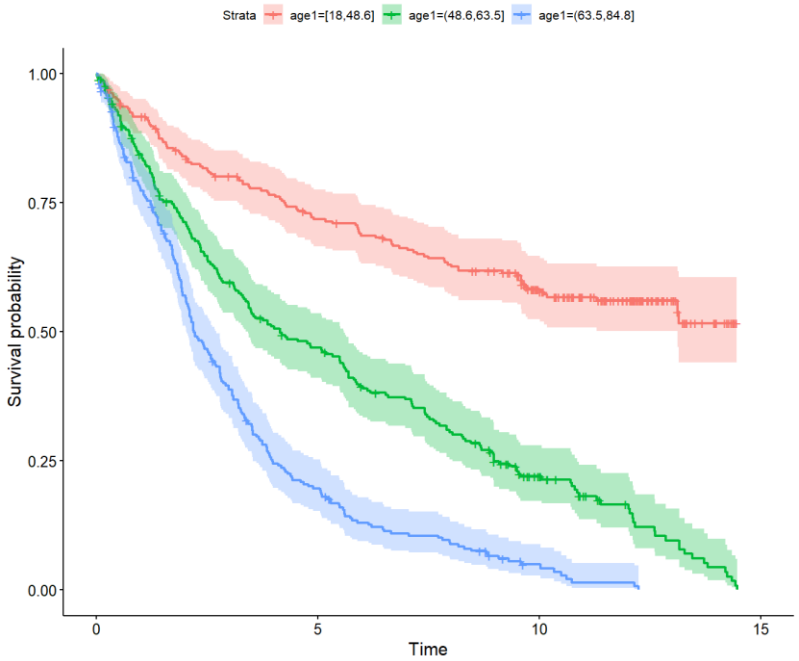
# Project: Survival analysis on diabetic patients

The following code was used to setup the data for survival analysis with death as the event.

```
#Death is the event so censor should be death = 1, everything else =
0.

table(dialysis$stat1)

#576 died

#223 not

dialysis.d <- subset(dialysis)

#subset

dialysis.d$censor <- as.integer(ifelse(dialysis.d$stat1 == 0, 1, 0))

table(dialysis.d$censor)

#correct, 576 registered as dead, rest are censored

#1 value used as death
```

After separating age into three groups of equal size, the following KM curve was produced.




From what can be observed, overall survival probability is clearly higher for the youngest age group (18-49) and worst for the older age groups. Survival probability has a severe decline for age groups 3 within the first five years while age group 2 has an almost constant decline in survival probability and finally age group 3 has a slight decline throughout the entire study.

Printing the survival model in R provides the median survival time of each group along with their respective confidence intervals.

```
print(death.fit)

                    n  events median 0.95LCL 0.95UCL
age1=[18,48.6]    267     112     NA   13.12      NA
age1=(48.6,63.5] 266     218   4.15    3.42    5.59
age1=(63.5,84.8] 266     246   2.20    2.04    2.71
```

The median survival cannot be obtained for the lowest age group as less than half of the patients experienced the event, thus group 1 never declines under 0.5% survival probability.

For each age group, the following survival probabilities at 10 years were determined.

| | Survival Probability | 95% LCI | 95% LCI |
|---|---|---|---|
| Age [18 – 48.6] | 0.58 | 0.52 | 0.65 |
| Age [48.6 – 63.5] | 0.22 | 0.17 | 0.28 |
| Age [63.5 – 84.8] | 0.05 | 0.03 | 0.09 |

Testing age categorical variable via logrank and trend.

```
death.fit2 <- survdiff(y.surv ~ age1, data = dialysis.d) #logrank

death.fit2
```

```
Call:
survdiff(formula = y.surv ~ age1, data = dialysis.d)

                    N Observed Expected (O-E)^2/E (O-E)^2/V
age1=[18,48.6]    267      112      272     93.92     189.2
age1=(48.6,63.5] 266      218      183      6.55       9.7
age1=(63.5,84.8] 266      246      121    129.46     174.9

 Chisq= 250  on 2 degrees of freedom, p= <2e-16
```

```
death.fit3 <- survfit(y.surv ~ age1, data = dialysis.d)

surv_pvalue(death.fit3, test.for.trend = TRUE) #trend test
```

From the results we can determine that there is a significant difference between the age categories alongside a monotonic trend of survival.

Testing for evidence of effect using wald test.

```
surv.model1 <- coxph(y.surv ~ age1, data = dialysis.d)

summary(surv.model1)
```

```
coxph(formula = y.surv ~ age1, data = dialysis.d)

  n= 799, number of events= 576

                   coef exp(coef) se(coef)      z Pr(>|z|)
age1(48.6,63.5] 1.1349    3.1109   0.1177  9.642   <2e-16 ***
age1(63.5,84.8] 1.7715    5.8799   0.1206 14.688   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                exp(coef) exp(-coef) lower .95 upper .95
age1(48.6,63.5]     3.111     0.3215     2.470     3.918
age1(63.5,84.8]     5.880     0.1701     4.642     7.448

Concordance= 0.654  (se = 0.011 )
Likelihood ratio test= 249.3  on 2 df,   p=<2e-16
Wald test            = 215.9  on 2 df,   p=<2e-16
Score (logrank) test = 249.9  on 2 df,   p=<2e-16
```

Wald test provides a value of <0.001, therefore there is evidence of a significant effect of age group on survival and risk of death.

Using covariates age and diabetes, a cox model was fitted with death as the event.

| Characteristic | HR[1] | 95% CI[1] | p-value |
|---|---|---|---|
| **Age** | 1.06 | 1.05, 1.06 | <0.001 |
| **Diabetes** | | | <0.001 |
| No | — | — | |
| Yes | 2.28 | 1.91, 2.71 | |

[1]HR = Hazard Ratio, CI = Confidence Interval

Both age and diabetes are significant covariates, where each increase of 1 year of age increases the risk of death by 6% and if a patient has diabetes, then they have 2.28 times the risk of death than someone without diabetes.

**Examining transplant as event instead of death**

Using transplant as the event instead and producing a table of univariate analysis with covariates: gender, diabetes, age and age (categorised).

| Characteristic | N | HR[1] | 95% CI[1] | p-value |
|---|---|---|---|---|
| **Age** | 799 | 0.98 | 0.97, 1.00 | 0.030 |
| **Gender** | 799 | | | 0.8 |
| Female | | — | — | |
| Male | | 0.95 | 0.66, 1.35 | |
| **Diabetes** | 799 | | | <0.001 |
| No | | — | — | |
| Yes | | 2.84 | 1.80, 4.47 | |
| **Age (Categorised)** | 799 | | | 0.001 |
| 18 - 49 | | — | — | |
| 49 - 64 | | 0.40 | 0.24, 0.69 | |
| 64 - 85 | | 0.89 | 0.28, 2.83 | |

[1]HR = Hazard Ratio, CI = Confidence Interval

After performing univariate analysis, it can be noted that only variables, Age, Diabetes and Age (Categorised) were significant ($p < 0.05$) when tested. Therefore, gender was excluded when building a multivariable model since it had a p-value of 0.8.

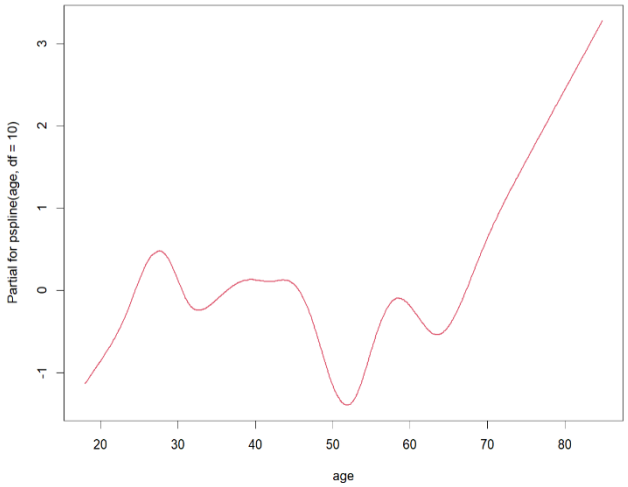
The continuous variable 'Age' was then centred using the mean and KM curves were generated for the categorical variables (even gender), where there appeared to be a distinct difference in survival when comparing diabetic patients and non-diabetic patients. 'Age (categorical)' had overlapping curves which indicate non-proportional hazards in terms of transplant. Therefore, 'Age' will be used instead of 'Age (categorical)' in the final model.

'Age' was then checked for linearity in a model with diabetes and it was found that the linear component was inadequate, but the non-linear component was significant, thus it required transformation. After testing polynomials, fractional polynomials and splines, it was found that splines with 10 degrees of freedom was the best choice and had the lowest AIC. Any other nonlinear models aside from spline models had a difference of over 10 AIC and thus a more interpretable model was not available.

Checking diagnostics for the model shows that proportional hazards assumption is not violated (Global $p = 0.35$). Examining the DFBETAs showed a concerning observation with 391 and 570, examining these two observations further shows that there are no unusual readings in terms of age or diabetes so they will not be removed.

| Characteristic | HR[1] | 95% CI[1] | p-value |
|---|---|---|---|
| **Diabetes** | | | <0.001 |
| No | — | — | |
| Yes | 2.85 | 1.79, 4.54 | |

[1]HR = Hazard Ratio, CI = Confidence Interval




To summarise the results of the multivariate model, it was found that both diabetes and age were significant predictors of time to transplant. Where patients with diabetes have 2.85 times the risk of transplant compared to patients without. In terms of age, the risk of transplant varies in a non-linear manner throughout the range, where patients initially have a linear increase in risk of transplant until around 28 years where the risk drops and plateaus before reaching the age where there is the lowest risk of transplant at around age 52. The risk then raises again until 60 and drops slightly before a linear increase in risk is clearly observable where patients around the age of 80 have the highest risk of transplant.

**APPENDIX (R code)**

```
dialysis <- read.csv("C:/Users/Kieran/Desktop/Stats
Datasets/Semester 2 2022/Regression/Regression Assignment
2/dialysis.csv", stringsAsFactors=TRUE)



library(survival)

library(survminer)

library(gtools)

library(gtsummary)

library(dplyr)

library(tidyverse)

library(flextable)

library(mfp)


#Q1i

#Death is the event so censor should be death = 1, everything else
= 0.

table(dialysis$stat1)

#576 died

#223 not

dialysis.d <- subset(dialysis)

#subset

dialysis.d$censor <- as.integer(ifelse(dialysis.d$stat1 == 0, 1,
0))

table(dialysis.d$censor)

#correct, 576 registered as dead, rest are censored

#1 value used as death


#q1ii

dialysis.d$age1 <- quantcut(dialysis.d$age, 3)

table(dialysis.d$age1)

#categorised into groups of 3
```

```
dialysis.d$y.surv <- Surv(dialysis.d$yrstotal, dialysis.d$censor)


death.fit <- survfit(y.surv ~ age1, data = dialysis.d)

ggsurvplot(death.fit, conf.int = TRUE)

#Survival.prob is clearly higher for ages 18-48, conversely prob is
worst for old ages.

#prob for older ages has a severed decline within the first 5 years
before being less severe.

#prob for young 18-48 is a gradual decline.


#q1iii

print(death.fit)

#There is not median as the lowest age category never breachs 0.5
survival probability.

#Less than 133 events


#q1iv

summary(death.fit, times=c(10))

#low cat: 0.58 (0.52:0.65)

#mid.cat: 0.22 (0.17:0.28)

#high.cat: 0.049 (0.028: 0.089)


#q1v

death.fit2 <- survdiff(y.surv ~ age1, data = dialysis.d) #logrank

death.fit2

#significant


death.fit3 <- survfit(y.surv ~ age1, data = dialysis.d)

surv_pvalue(death.fit3, test.for.trend = TRUE) #trend

#significant monotonic relationship with ordered categories [linear
castegory test]
```

```
#q1vi Cox model, age group, test for evidence of effect of age
using wald.

surv.model1 <- coxph(y.surv ~ age1, data = dialysis.d)

summary(surv.model1)


tbl_regression(surv.model1) %>%

  add_global_p()

#Using Wald test, there is evidence that age(Category) has a
significant effect


#q1vii COX age and diabetes

#factor diabetes

dialysis.d$diabetesf <- factor(dialysis.d$diabetes)


surv.model2 <- coxph(y.surv ~ age + diabetesf, data = dialysis.d)

tbl_regression(surv.model2) %>%

  add_global_p()


levels(dialysis.d$diabetesf) <- c("No", "Yes")



tbl_regression(surv.model2, exponentiate = TRUE,

               label = list (age ~ "Age",

                             diabetesf ~ "Diabetes")) %>%

  add_global_p() %>%

  bold_labels() %>%

  as_flex_table()



#per 1 year of age increases risk of death by 6%

#If diabetes is present then patient has 2.28 times the risk of
death than someone without diabetes
```

```
#q2i

#Transplant is now the event so censor should be Dialysis = 1,
everything else = 0.

dialysis.t <- subset(dialysis)

#128 dialysis cases


dialysis.t$censor <- as.integer(ifelse(dialysis.t$stat1 == 2, 1,
0))

table(dialysis.t$censor)

# age cat

dialysis.t$age.cat <- quantcut(dialysis.t$age, 3)

#new y.surv

dialysis.t$y.surv <- Surv(dialysis.t$yrstotal, dialysis.t$censor)



#data mutation, only select variables I want and factoring
variables

#only need y.surv, age, gender, diabetes, age.cat

dialysis.t$gender <- factor(dialysis.t$gender)

dialysis.t$diabetes <- factor(dialysis.t$diabetes)


levels(dialysis.t$age.cat) <- c("18 - 49", "49 - 64", "64 - 85")


dialysis.t %>%
  select(y.surv, age, gender, diabetes, age.cat) %>%
  mutate(
    gender = case_when(gender == 0 ~ "Female",
                       gender == 1 ~ "Male"),
    diabetes = case_when(diabetes == 0 ~ "No",
                         diabetes == 1 ~ "Yes")
  ) -> dialysis.t2
```

```
dialysis.t2 %>%
  tbl_uvregression(
    method = coxph,
    y = y.surv,
    exponentiate = TRUE,
    label = list(age ~ "Age",
                 gender ~ "Gender",
                 diabetes ~ "Diabetes",
                 age.cat ~ "Age (Categorised)")) %>%
  bold_labels() %>%
  add_global_p() %>%
  as_flex_table()



#q2ii
colSums(is.na(dialysis.t2[1:5]))
#Just checking for missing values but no missing values so AIC is a
valid model validation method


#AGE_cent
hist(dialysis.t2$age)
dialysis.t2$age_cent <- dialysis.t2$age - 54.05907
#----


#KM plot
transplant.fit <- survfit(y.surv ~ age.cat, data = dialysis.t2)
ggsurvplot(transplant.fit, conf.int = TRUE)


transplant.fit <- survfit(y.surv ~ diabetes, data = dialysis.t2)
ggsurvplot(transplant.fit, conf.int = TRUE)
```

```
transplant.fit <- survfit(y.surv ~ gender, data = dialysis.t2)
ggsurvplot(transplant.fit, conf.int = TRUE)

print(transplant.fit)



model1.1 <- coxph(y.surv ~ age + gender + diabetes + age.cat, data
= dialysis.t2)
car::Anova(model1.1)
modelph <- cox.zph(model1.1)
modelph
#Age categorical confirmed to be breaching PH along with continuous



tbl_regression(model1.1, exponentiate = TRUE,
               label = list (age ~ "Age",
                             diabetes ~ "Diabetes")) %>%
  add_global_p() %>%
  bold_labels() %>%
  as_flex_table()

model1.3 <- coxph(y.surv ~ age_cent + diabetes, data = dialysis.t2)
model1.3

model1.4 <- coxph(y.surv ~ diabetes + age.cat, data = dialysis.t2)

AIC(model1.1)
AIC(model1.2)
AIC(model1.3)
AIC(model1.4)
```

```
coxph(y.surv ~ age + gender + diabetes + age.cat, data =
dialysis.t2) %>%

AIC()




#AIC for age


modelage.1 <- coxph(y.surv ~ pspline(age, df=4) + diabetes, data =
dialysis.t2)

modelage.2 <- coxph(y.surv ~ pspline(age, df=5) + diabetes, data =
dialysis.t2)

modelage.3 <- coxph(y.surv ~ pspline(age, df=6) + diabetes, data =
dialysis.t2)

modelage.4 <- coxph(y.surv ~ pspline(age, df=7) + diabetes, data =
dialysis.t2)

modelage.5 <- coxph(y.surv ~ pspline(age, df=8) + diabetes, data =
dialysis.t2)

modelage.6 <- coxph(y.surv ~ pspline(age, df=9) + diabetes, data =
dialysis.t2)

modelage.7 <- coxph(y.surv ~ pspline(age, df=10) + diabetes, data =
dialysis.t2) #best spline

modelage.8 <- coxph(y.surv ~ pspline(age, df=11) + diabetes, data =
dialysis.t2)


termplot(modelage.1)

hist(dialysis.t2$age)

hist(dialysis.t2$agelog)

hist(dialysis.t2$agesqrt)


AIC(modelage.1)

AIC(modelage.2)

AIC(modelage.3)
```

```
termplot(modelage.3, term=1)

AIC(modelage.4)

termplot(modelage.4, term=1)

AIC(modelage.5)

AIC(modelage.6)

AIC(modelage.7) #best model ------------------------------------------
----------------------

AIC(modelage.8)


dialysis.t2$agelog <- log(dialysis.t2$age)

dialysis.t2$agesqrt <- sqrt(dialysis.t2$age)


modelage.9 <- coxph(y.surv ~ dialysis.t2$agelog + diabetes, data =
dialysis.t2)

modelage.10 <- coxph(y.surv ~ dialysis.t2$agesqrt + diabetes, data
= dialysis.t2)


AIC(modelage.9)

AIC(modelage.10)


dialysis.t2$age_centsquared <- dialysis.t2$age_cent^2

dialysis.t2$age_centsquared2 <-
dialysis.t2$age_centsquared*dialysis.t2$age_centsquared


modelage.squared <- coxph(y.surv ~ dialysis.t2$age_cent +
age_centsquared + diabetes, data = dialysis.t2)

modelage.squared2 <- coxph(y.surv ~ dialysis.t2$age_cent +
age_centsquared + age_centsquared2 + diabetes, data = dialysis.t2)


AIC(modelage.squared)

AIC(modelage.squared2)


#fractional polynomial for age
```

```
modelage.fp <- coxph(y.surv ~ fp(age) + diabetes, data =
dialysis.t2)

modelage.fp

AIC(modelage.fp)

modelph <- cox.zph(modelage.7)

modelph



#Using best spline ----------------------------------------------
----------------

finalmodel.t <- coxph(y.surv ~ pspline(age, df=10) + diabetes, data
= dialysis.t2)

finalmodel.t

f.plot <- termplot(finalmodel.t, term=1, se=TRUE)

f.plot

termplot(finalmodel.t, term=1)



tbl_regression(finalmodel.t, exponentiate = TRUE,

               label = list (diabetes ~ "Diabetes")) %>%

  add_global_p() %>%

  bold_labels() %>%

  as_flex_table()


#diagnostics

modelph <- cox.zph(finalmodel.t)

modelph



finalmodel.res<- residuals(finalmodel.t, type="dfbeta")

n <- nrow(finalmodel.res)

## reshape into long format for plotting
```

```
finalmodel.res2 <- data.frame(subject=rep(rownames(finalmodel.res),
2),
                          DFBETA=c(finalmodel.res[,1],
finalmodel.res[,2]),
                          variable=c(rep("age",n ),rep("Yes",n )))
library(car)
Boxplot(DFBETA ~ variable, data=finalmodel.res2,
        id=list(labels=finalmodel.res2$subject, cex=0.5),
        cex.axis = 0.7)


dialysis.t2[c(391, 570), c("y.surv", "age", "diabetes")]
```