01\_homework1

Randy

1/29/2021

# BIOS7721 Homework1

## Question1 longitudinal analysis

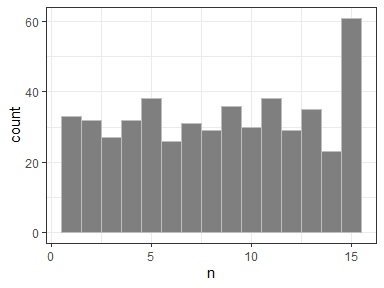
### a. the number of measurements varies

* calculate each subject measurement number
* distribution of aortic gradient
* why consider square root transformation
* create sqrt.aort.gard column for transformation

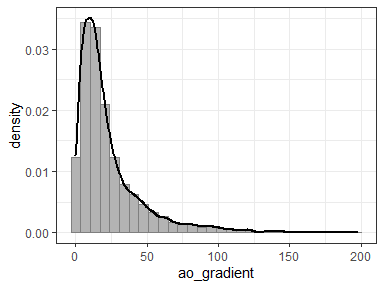
The dataset is not well balanced for each individual.  
The measurement times are calculated and presented in *plot\_count*.  
The distribution of aortic gradient is presented in *plot\_hist\_ag*.  
According to the *plot\_hist\_ag*, the distribution of aortic gradient is highly right-skewed.  
To balance the long right tail, the square root transformation is applied. After the transformation, the sqrt aortic gradient looks more similar to normal distribution.

aort <- here::here("aort.csv") %>%   
 read\_csv() %>%  
 janitor::clean\_names()  
  
## to count each id's measurement  
aort\_count <- aort %>%   
 group\_by(id) %>%  
 count()   
# aort\_count  
  
## to get the frequency of measurement  
aort\_recount <-  
 aort\_count %>%  
 group\_by(n) %>%  
 count()  
# aort\_recount

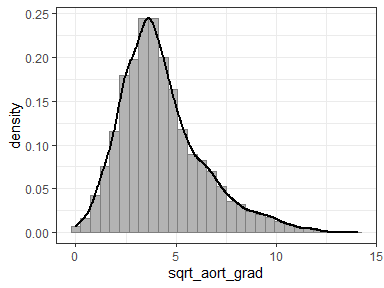
## plot for measurement frequency  
plot\_count <-   
 aort\_count %>%  
 ggplot(aes(n, fill = "aort")) +  
 geom\_histogram(binwidth = 1,  
 fill = "grey50",   
 color = "grey") +  
 theme(legend.position = "none") +  
 theme\_bw()  
plot\_count



## plot for the ag distribution  
plot\_hist\_ag <- aort %>%  
 ggplot(aes(ao\_gradient)) +  
 geom\_histogram(aes(y = ..density..),  
 fill = "grey70",  
 color = "grey50") +  
 geom\_density(alpha = 0.1,   
 size = 1) +  
 theme(legend.position="none") +  
 theme\_bw()  
plot\_hist\_ag



## to add extra sqrt\_aort\_grad  
aort1 <- aort %>%  
 mutate(sqrt\_aort\_grad = sqrt(ao\_gradient),  
 event = as.factor(event),  
 id= as.factor(id))  
  
## after sqrt transformation  
plot\_hist\_sag <- aort1 %>%  
 ggplot(aes(sqrt\_aort\_grad)) +  
 geom\_histogram(aes(y = ..density..),  
 fill = "grey70",  
 color = "grey50") +  
 geom\_density(alpha = 0.1,   
 size = 1) +  
 theme(legend.position="none") +  
 theme\_bw()  
plot\_hist\_sag



head(aort1, 5)

## # A tibble: 5 x 9  
## id ao\_gradient time ev\_time event type\_op sex age sqrt\_aort\_grad  
## <fct> <dbl> <dbl> <dbl> <fct> <chr> <chr> <dbl> <dbl>  
## 1 1 4.32 0 1.86 0 SI Male 43.7 2.08  
## 2 1 3.65 0.750 1.86 0 SI Male 43.7 1.91  
## 3 2 8.25 0 12.6 0 SI Male 69.1 2.87  
## 4 2 6.43 3.23 12.6 0 SI Male 69.1 2.54  
## 5 2 27.4 3.82 12.6 0 SI Male 69.1 5.24

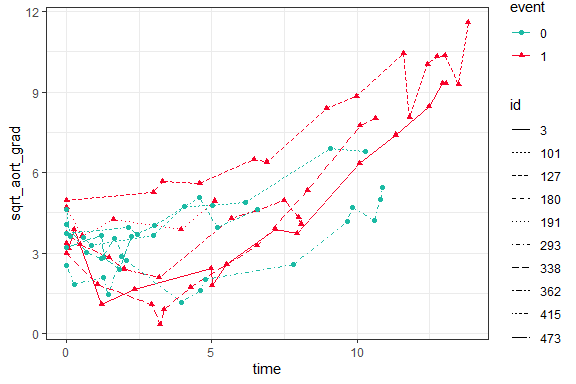
### b. subset 5 subjects

* for with or without events
* plot and describe observed trends

The trend are affected by the samples chosen. Based on these ten samples, the subjects without events have shorter follow-up time; also some of their aortic gradient level increases slower over the time, but the others aortic gradient trends have no difference with the subjects suffered events during follow-up. This indicates potential informative censor or missing not at random cases. Both groups time trend can be approximated through linear or quadratic trend.

set.seed(55)  
aort\_live <- aort1 %>%  
 group\_by(id) %>%  
 filter(event == 0) %>%  
 ## to nest the dataset   
 ## in one cell  
 nest() %>%  
 as.tibble() %>%  
 sample\_n(size = 5) %>%  
 unnest()  
  
aort\_dead <- aort1 %>%  
 group\_by(id) %>%  
 filter(event == 1) %>%  
 ## to nest the dataset   
 ## in one cell  
 nest() %>%  
 as.tibble() %>%  
 sample\_n(size = 5) %>%  
 unnest()

plot\_sample <-   
 rbind(aort\_dead, aort\_live) %>%  
 ggplot(aes(x = time,  
 y = sqrt\_aort\_grad,  
 group = id,  
 linetype = id,  
 color = event)) +  
 geom\_line() +  
 geom\_point(aes(shape = event)) +  
 theme\_bw() +  
 scale\_color\_manual(values = c("#1cbaa4", "#f7022a"))  
  
plot\_sample



### c. fit a random intercept model

* sqrt\_aort\_grad as outcome
* linear effect of time
* mean aortic gradient over time?
* change vary be surgery?
* variation of baseline between subject?
* hypothesis test
* interpret the coefficient

According to the random intercept model (model1), the time has a very highly significant effect on the subject’s aortic gradient level (p << 0.001). Hence, there is an evidence for the mean aortic fradient changing over time.  
According to the model adjusted for operation type (model2), there is a very highly significant effect on the subject’s aortic gradient level over different operation type (p << 0.001); however the time effect on different operation type are not significant (p = 0.67 > 0.05).  
There is a variability for each subject aortic gradient level.

model1 <- lme(sqrt\_aort\_grad ~ time,  
 random = ~1 | id,  
 data = aort1)   
model2 <- lme(sqrt\_aort\_grad ~ time \* type\_op,  
 random = ~1 | id,  
 data = aort1)   
  
tidy1 <- broom.mixed::tidy(model1)  
tidy2 <- broom.mixed::tidy(model2)  
  
glance1 <- broom.mixed::glance(model1)  
glance2 <- broom.mixed::glance(model2)  
# augment1 <- broom.mixed::augment(model1)  
# augment2 <- broom.mixed::augment(model2)

## effects for model1, model2  
tidy1

## # A tibble: 4 x 8  
## effect group term estimate std.error df statistic p.value  
## <chr> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 fixed fixed (Intercept) 2.61 0.0566 3682 46.1 0  
## 2 fixed fixed time 0.391 0.00462 3682 84.6 0  
## 3 ran\_pars id sd\_(Intercept) 1.13 NA NA NA NA  
## 4 ran\_pars Residual sd\_Observation 1.00 NA NA NA NA

tidy2

## # A tibble: 6 x 8  
## effect group term estimate std.error df statistic p.value  
## <chr> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 fixed fixed (Intercept) 2.23 0.0786 3681 28.4 2.83e-160  
## 2 fixed fixed time 0.393 0.00644 3681 61.0 0.   
## 3 fixed fixed type\_opSI 0.721 0.109 498 6.62 9.59e- 11  
## 4 fixed fixed time:type\_opSI -0.00383 0.00923 3681 -0.415 6.78e- 1  
## 5 ran\_pars id sd\_(Intercept) 1.07 NA NA NA NA   
## 6 ran\_pars Residual sd\_Observation 1.00 NA NA NA NA

To see the subject individual aortic gradient variability to baseline level, the baseline aortic gradient level is adjusted in model3. According to model3, baseline aortic gradient level can significantly affect the subject’s future aortic gradient level. These results indicate that there is a variation for baseline aortic gradient. After adjustment on the baseline level, the model performs better.

aort2 <- aort1 %>%   
 mutate(baseline = ifelse(time == 0, ao\_gradient, NA)) %>%  
 fill(baseline) %>%  
 mutate(basesqrt = sqrt(baseline))  
  
model3 <- lme(sqrt\_aort\_grad ~ basesqrt + time \* type\_op,  
 random = ~ 1 | id,  
 data = aort2)  
  
tidyf3 <- broom.mixed::tidy(model3)  
tidyr3 <- broom.mixed::tidy(model3, effects = "ran\_pars", conf.int = TRUE)  
glance3 <- broom.mixed::glance(model3)

tidyf3

## # A tibble: 7 x 8  
## effect group term estimate std.error df statistic p.value  
## <chr> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 fixed fixed (Intercept) 0.0860 0.133 3681 0.648 5.17e- 1  
## 2 fixed fixed basesqrt 0.762 0.0415 497 18.4 6.39e-58  
## 3 fixed fixed time 0.393 0.00635 3681 61.9 0.   
## 4 fixed fixed type\_opSI 0.140 0.0932 497 1.50 1.34e- 1  
## 5 fixed fixed time:type\_opSI -0.00490 0.00909 3681 -0.539 5.90e- 1  
## 6 ran\_pars id sd\_(Intercept) 0.797 NA NA NA NA   
## 7 ran\_pars Residual sd\_Observation 0.998 NA NA NA NA

tidyr3

## # A tibble: 2 x 6  
## effect group term estimate conf.low conf.high  
## <chr> <chr> <chr> <dbl> <dbl> <dbl>  
## 1 ran\_pars id sd\_(Intercept) 0.797 0.740 0.859  
## 2 ran\_pars Residual sd\_Observation 0.998 NA NA

As seen for the criteria comparison cross the three models, model3, adjusting both operation type and the baseline aortic gradients, performs best with lowest AIC (12748) and BIC (12793).

Also based on log-likelihood, the chi-squre and pvalue calculated among nested models. The results indicate that the adjustment for both operation type and the baseline aortic gradients can significantly improve the inference model.

For model3, there is a significant time trend for subject’s aortic gradient level (p << 0.001). On average, in population level, the subject’s sqrt aortic gradient level increase 0.39 (95% CI: 0.38, 0.41) unit in each year. Also the operation type does not significantly affect the patient’s aortic gradient level (p = 0.13), and there is no significant time:operation interaction effect (p = 0.59). Intriguingly, the subject’s baseline aortic gradient level can also affect the future (p << 0.01). On average, in population level, the patient, with higher baseline aortic gradient, The increasing each unit of baseline sqrt aortic gradient will increase 0.76 (95% CI: 0.68, 0.84) unit in future sqrt aortic gradient level accordingly. As random effects, there is a large standard deviation on sqrt aortic gradient level (sd = 0.78) on individual level. Also the random effects residual standard deviation is pretty large too. These indicate the individual level variation could not be ignored but treated as random effects individually.

rbind(glance1,   
 glance2,  
 glance3) %>%  
 rownames\_to\_column("model")

## # A tibble: 3 x 5  
## model sigma logLik AIC BIC  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 1 1.00 -6513. 13033. 13058.  
## 2 2 1.00 -6495. 13002. 13040.  
## 3 3 0.998 -6367. 12749. 12793.

#' test\_lrt() is to calculate the lrt pvalue  
#'  
#' @param mod0 is the first model  
#' @param mod1 is the second model  
#' @param df can be add manually  
#' @return pvalue is the lrt pvalue  
#' @examples  
#' test\_lrt(model1, model2)  
test\_lrt <- function(mod0, mod1, ...){  
 A <- logLik(mod0)   
 B <- logLik(mod1)   
 D <- -2 \* (as.numeric(A) - as.numeric(B))  
 df = abs(attributes(A)$df - attributes(B)$df)  
 pvalue <- pchisq(D, df = df,   
 lower.tail = FALSE)  
 return(pvalue)  
}  
  
test\_lrt(model1, model2, 1)

## [1] 2.930505e-08

test\_lrt(model2, model3)

## [1] 1.646453e-57

### d. extend to random slope model

* how many more parameters
* outcome change over time between subject?

Two more parameters are estimated, the estimate for random slope and its correlation matrix. The random slope on time has standard deviation 0.18, with mean set as zero. In this case, the variability of random time slope effects among subjects cannot be ignored.

After addition of random slope term the AIC decreased 884 to 11864; this indicates the variability on random time slope term for each individual.

model4 <- lme(sqrt\_aort\_grad ~ basesqrt + time \* type\_op,  
 random = ~ time + 1 | id,   
 data = aort2)   
tidy4 <- broom.mixed::tidy(model4)  
glance4 <- broom.mixed::glance(model4)  
augment4 <- broom.mixed::augment(model4)

## the estimators for model4  
tidy4

## # A tibble: 9 x 8  
## effect group term estimate std.error df statistic p.value  
## <chr> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 fixed fixed (Intercept) 0.134 0.104 3681 1.28 1.99e- 1  
## 2 fixed fixed basesqrt 0.765 0.0329 497 23.2 2.24e- 81  
## 3 fixed fixed time 0.354 0.0152 3681 23.4 1.40e-112  
## 4 fixed fixed type\_opSI 0.199 0.0728 497 2.74 6.35e- 3  
## 5 fixed fixed time:type\_opSI -0.0376 0.0214 3681 -1.76 7.83e- 2  
## 6 ran\_pars id sd\_(Intercept) 0.566 NA NA NA NA   
## 7 ran\_pars id cor\_time.(Inte~ -0.159 NA NA NA NA   
## 8 ran\_pars id sd\_time 0.179 NA NA NA NA   
## 9 ran\_pars Residu~ sd\_Observation 0.842 NA NA NA NA

rbind(glance1,  
 glance2,  
 glance3,  
 glance4) %>%  
 tibble() %>%  
 rownames\_to\_column("model")

## # A tibble: 4 x 5  
## model sigma logLik AIC BIC  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 1 1.00 -6513. 13033. 13058.  
## 2 2 1.00 -6495. 13002. 13040.  
## 3 3 0.998 -6367. 12749. 12793.  
## 4 4 0.842 -5923. 11864. 11921.

### e. explore splines in fixed and random effects

* comments on results

Because the earlier models indicates the time:operation interaction term did not contribute to the model performance, the interaction term is removed from model6. After remove the interaction, there is no significant improvement for model6 (AIC = 9893) than model5 (AIC = 9899).

The B-spline adds more flexibility to the model, so the over all fitting performs much better (with much lower AIC). The time trend and operation type can all significantly affect the aortic gradient level; but the time:operation interaction terms are still not significant, which is consistent for model5 and model6. The overall results are pretty consistent with earlier models.

model5 <- lme(sqrt\_aort\_grad ~ bs(time) \* type\_op,  
 random = list(id = pdDiag(form = ~ bs(time))),  
 ## if the entire matrix used  
 ## then there might be a numerical problem  
 data = aort2)  
  
model6 <- lme(sqrt\_aort\_grad ~ bs(time) + type\_op,  
 random = list(id = pdDiag(form = ~ bs(time))),  
 ## if the entire matrix used  
 ## then there might be a numerical problem  
 data = aort2)  
  
tidy5 <- broom.mixed::tidy(model5, effects = "fixed")  
glance5 <- broom.mixed::glance(model5)  
augment5 <- broom.mixed::augment(model5)  
tidy6 <- broom.mixed::tidy(model6, effects = "fixed")  
glance6 <- broom.mixed::glance(model6)  
augment6 <- broom.mixed::augment(model6)

tidy5

## # A tibble: 8 x 6  
## term estimate std.error df statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) 2.89 0.0571 3677 50.6 0.   
## 2 bs(time)1 -0.594 0.203 3677 -2.92 3.48e- 3  
## 3 bs(time)2 2.65 0.292 3677 9.07 1.94e- 19  
## 4 bs(time)3 8.24 0.329 3677 25.0 1.10e-127  
## 5 type\_opSI 0.688 0.0789 498 8.73 3.99e- 17  
## 6 bs(time)1:type\_opSI 0.00299 0.285 3677 0.0105 9.92e- 1  
## 7 bs(time)2:type\_opSI -0.0244 0.415 3677 -0.0588 9.53e- 1  
## 8 bs(time)3:type\_opSI -0.240 0.475 3677 -0.505 6.14e- 1

tidy6

## # A tibble: 5 x 6  
## term estimate std.error df statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) 2.89 0.0561 3680 51.5 0.   
## 2 bs(time)1 -0.594 0.142 3680 -4.18 3.01e- 5  
## 3 bs(time)2 2.64 0.207 3680 12.7 2.66e- 36  
## 4 bs(time)3 8.12 0.237 3680 34.3 3.14e-224  
## 5 type\_opSI 0.686 0.0762 498 9.00 4.72e- 18

rbind(glance1,  
 glance2,  
 glance3,  
 glance4,  
 glance5,  
 glance6) %>%  
 tibble() %>%  
 rownames\_to\_column("model")

## # A tibble: 6 x 5  
## model sigma logLik AIC BIC  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 1 1.00 -6513. 13033. 13058.  
## 2 2 1.00 -6495. 13002. 13040.  
## 3 3 0.998 -6367. 12749. 12793.  
## 4 4 0.842 -5923. 11864. 11921.  
## 5 5 0.575 -4937. 9900. 9982.  
## 6 6 0.575 -4937. 9893. 9957.

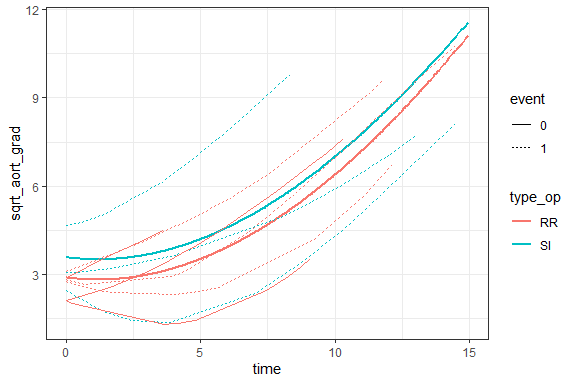
### f. plot the population trajectory

* plot the predict 10 patients trajectory

Over all the trajectory can be approximated through quadratic or cubic function as increasing pattern.  
In population level, the SI operation has a higher aortic gradient level than the RR operation.  
There is a really high variability on the random effect on subject level. Also the subjects did not suffer events have shorter follow-up time, even the trends seem similar to the subjects with event. There is a chance that the missing mechanism is not random; without further analysis on the missing mechanism or data implement, the results might be biased.

set.seed(5555)  
augment5 <- broom.mixed::augment(model5) %>%   
 as.data.frame()  
  
## use level = 0 for the population  
augment5\_sub <-  
 augment5 %>%  
 group\_by(id) %>%  
 nest() %>%  
 tibble() %>%  
 sample\_n(size = 10,  
 replace = FALSE) %>%  
 unnest() %>%  
 tibble()

plot\_model5 <-   
 ggplot(augment5) +  
 geom\_line(aes(x = time,   
 y = .fixed,  
 group = type\_op,  
 color = type\_op),  
 linetype = "solid",  
 size = 1) +  
 geom\_line(data = augment5\_sub,  
 aes(x = time,  
 y = .fitted,  
 group = id,  
 linetype = event,  
 color = type\_op)) +  
 xlab("time") +  
 ylab("sqrt\_aort\_grad") +  
 theme\_bw()  
  
plot\_model5



### g. variance-covariance matrix for patient

* correlation matrix for patient
* from model2, model3, and model4

For model2 the covariance and correlation matrices are positive definite, the correlation keeps the same cross the time.  
For model3 the covariance and correlation matrices are still positive definite and symmetric, but the correlation decreases over time.  
For model5 the covariance and correlation matrices are still positive definite and symmetric, but there is no obvious pattern over time.

#' get\_cov\_cor() is for covariance and correlation  
#'  
#' @param object a lm or lmm object  
#' @param id the subjec interested with  
#' @param type options: "margin", "residual", and "condition"  
#' @return mcov(marginal covariance matrix)   
#' @return cor(correlation matrix based on mcov)  
#' @examples  
#' get\_cov\_cor(model1, 5, "condition")  
get\_cov\_cor <-   
 function(object, id, type = "margin") {  
 mcov <- nlme::getVarCov(  
 obj = object,  
 individual = id,  
 type = type)  
 cor <- stats::cov2cor(mcov[[1]]) %>%  
 round(2)  
   
 mcov <- mcov[["2"]] %>%   
 round(2)  
 return(list(mcov, cor))  
 }

get\_cov\_cor(model2, 2)

## [[1]]  
## 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15  
## 1 2.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15  
## 2 1.15 2.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15  
## 3 1.15 1.15 2.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15  
## 4 1.15 1.15 1.15 2.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15  
## 5 1.15 1.15 1.15 1.15 2.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15  
## 6 1.15 1.15 1.15 1.15 1.15 2.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15  
## 7 1.15 1.15 1.15 1.15 1.15 1.15 2.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15  
## 8 1.15 1.15 1.15 1.15 1.15 1.15 1.15 2.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15  
## 9 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 2.15 1.15 1.15 1.15 1.15 1.15 1.15  
## 10 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 2.15 1.15 1.15 1.15 1.15 1.15  
## 11 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 2.15 1.15 1.15 1.15 1.15  
## 12 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 2.15 1.15 1.15 1.15  
## 13 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 2.15 1.15 1.15  
## 14 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 2.15 1.15  
## 15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 1.15 2.15  
##   
## [[2]]  
## 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15  
## 1 1.00 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53  
## 2 0.53 1.00 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53  
## 3 0.53 0.53 1.00 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53  
## 4 0.53 0.53 0.53 1.00 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53  
## 5 0.53 0.53 0.53 0.53 1.00 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53  
## 6 0.53 0.53 0.53 0.53 0.53 1.00 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53  
## 7 0.53 0.53 0.53 0.53 0.53 0.53 1.00 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53  
## 8 0.53 0.53 0.53 0.53 0.53 0.53 0.53 1.00 0.53 0.53 0.53 0.53 0.53 0.53 0.53  
## 9 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 1.00 0.53 0.53 0.53 0.53 0.53 0.53  
## 10 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 1.00 0.53 0.53 0.53 0.53 0.53  
## 11 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 1.00 0.53 0.53 0.53 0.53  
## 12 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 1.00 0.53 0.53 0.53  
## 13 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 1.00 0.53 0.53  
## 14 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 1.00 0.53  
## 15 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.53 1.00

get\_cov\_cor(model4, 2)

## [[1]]  
## 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15  
## 1 1.03 0.27 0.26 0.24 0.23 0.23 0.21 0.21 0.19 0.18 0.17 0.16 0.15 0.14 0.13  
## 2 0.27 1.26 0.60 0.68 0.74 0.76 0.85 0.87 0.96 1.04 1.10 1.13 1.22 1.24 1.32  
## 3 0.26 0.60 1.38 0.76 0.83 0.86 0.97 1.00 1.10 1.20 1.27 1.31 1.41 1.44 1.54  
## 4 0.24 0.68 0.76 1.59 0.97 1.01 1.14 1.18 1.31 1.43 1.52 1.57 1.70 1.74 1.87  
## 5 0.23 0.74 0.83 0.97 1.78 1.11 1.27 1.31 1.47 1.60 1.71 1.77 1.92 1.96 2.11  
## 6 0.23 0.76 0.86 1.01 1.11 1.87 1.32 1.37 1.53 1.67 1.78 1.85 2.00 2.05 2.21  
## 7 0.21 0.85 0.97 1.14 1.27 1.32 2.23 1.58 1.77 1.95 2.08 2.15 2.34 2.40 2.59  
## 8 0.21 0.87 1.00 1.18 1.31 1.37 1.58 2.34 1.84 2.01 2.15 2.23 2.43 2.49 2.68  
## 9 0.19 0.96 1.10 1.31 1.47 1.53 1.77 1.84 2.78 2.28 2.44 2.53 2.76 2.83 3.05  
## 10 0.18 1.04 1.20 1.43 1.60 1.67 1.95 2.01 2.28 3.22 2.69 2.79 3.04 3.12 3.37  
## 11 0.17 1.10 1.27 1.52 1.71 1.78 2.08 2.15 2.44 2.69 3.59 2.99 3.26 3.35 3.62  
## 12 0.16 1.13 1.31 1.57 1.77 1.85 2.15 2.23 2.53 2.79 2.99 3.82 3.39 3.48 3.76  
## 13 0.15 1.22 1.41 1.70 1.92 2.00 2.34 2.43 2.76 3.04 3.26 3.39 4.41 3.80 4.11  
## 14 0.14 1.24 1.44 1.74 1.96 2.05 2.40 2.49 2.83 3.12 3.35 3.48 3.80 4.61 4.22  
## 15 0.13 1.32 1.54 1.87 2.11 2.21 2.59 2.68 3.05 3.37 3.62 3.76 4.11 4.22 5.28  
##   
## [[2]]  
## 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15  
## 1 1.00 0.24 0.22 0.19 0.17 0.17 0.14 0.13 0.11 0.10 0.09 0.08 0.07 0.07 0.05  
## 2 0.24 1.00 0.46 0.48 0.49 0.50 0.51 0.51 0.51 0.52 0.52 0.52 0.52 0.51 0.51  
## 3 0.22 0.46 1.00 0.52 0.53 0.54 0.55 0.56 0.56 0.57 0.57 0.57 0.57 0.57 0.57  
## 4 0.19 0.48 0.52 1.00 0.58 0.58 0.61 0.61 0.62 0.63 0.64 0.64 0.64 0.64 0.65  
## 5 0.17 0.49 0.53 0.58 1.00 0.61 0.64 0.64 0.66 0.67 0.68 0.68 0.68 0.69 0.69  
## 6 0.17 0.50 0.54 0.58 0.61 1.00 0.65 0.65 0.67 0.68 0.69 0.69 0.70 0.70 0.70  
## 7 0.14 0.51 0.55 0.61 0.64 0.65 1.00 0.69 0.71 0.73 0.73 0.74 0.75 0.75 0.75  
## 8 0.13 0.51 0.56 0.61 0.64 0.65 0.69 1.00 0.72 0.73 0.74 0.75 0.76 0.76 0.76  
## 9 0.11 0.51 0.56 0.62 0.66 0.67 0.71 0.72 1.00 0.76 0.77 0.78 0.79 0.79 0.80  
## 10 0.10 0.52 0.57 0.63 0.67 0.68 0.73 0.73 0.76 1.00 0.79 0.80 0.81 0.81 0.82  
## 11 0.09 0.52 0.57 0.64 0.68 0.69 0.73 0.74 0.77 0.79 1.00 0.81 0.82 0.82 0.83  
## 12 0.08 0.52 0.57 0.64 0.68 0.69 0.74 0.75 0.78 0.80 0.81 1.00 0.83 0.83 0.84  
## 13 0.07 0.52 0.57 0.64 0.68 0.70 0.75 0.76 0.79 0.81 0.82 0.83 1.00 0.84 0.85  
## 14 0.07 0.51 0.57 0.64 0.69 0.70 0.75 0.76 0.79 0.81 0.82 0.83 0.84 1.00 0.86  
## 15 0.05 0.51 0.57 0.65 0.69 0.70 0.75 0.76 0.80 0.82 0.83 0.84 0.85 0.86 1.00

get\_cov\_cor(model5, 2)

## [[1]]  
## 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15  
## 1 0.89 0.56 0.56 0.56 0.56 0.56 0.56 0.56 0.56 0.56 0.56 0.56 0.56 0.56 0.56  
## 2 0.56 1.72 1.46 1.54 1.58 1.58 1.58 1.58 1.53 1.46 1.39 1.35 1.23 1.19 1.05  
## 3 0.56 1.46 1.89 1.65 1.69 1.71 1.72 1.72 1.67 1.61 1.54 1.49 1.37 1.32 1.17  
## 4 0.56 1.54 1.65 2.10 1.84 1.85 1.89 1.90 1.87 1.82 1.76 1.71 1.59 1.54 1.37  
## 5 0.56 1.58 1.69 1.84 2.24 1.94 2.00 2.00 2.00 1.96 1.91 1.87 1.74 1.70 1.53  
## 6 0.56 1.58 1.71 1.85 1.94 2.29 2.03 2.04 2.05 2.02 1.97 1.93 1.81 1.76 1.59  
## 7 0.56 1.58 1.72 1.89 2.00 2.03 2.47 2.16 2.21 2.20 2.17 2.15 2.05 2.00 1.84  
## 8 0.56 1.58 1.72 1.90 2.00 2.04 2.16 2.52 2.24 2.24 2.22 2.20 2.10 2.06 1.90  
## 9 0.56 1.53 1.67 1.87 2.00 2.05 2.21 2.24 2.66 2.38 2.38 2.37 2.31 2.28 2.15  
## 10 0.56 1.46 1.61 1.82 1.96 2.02 2.20 2.24 2.38 2.78 2.48 2.49 2.47 2.45 2.35  
## 11 0.56 1.39 1.54 1.76 1.91 1.97 2.17 2.22 2.38 2.48 2.87 2.56 2.57 2.57 2.50  
## 12 0.56 1.35 1.49 1.71 1.87 1.93 2.15 2.20 2.37 2.49 2.56 2.92 2.62 2.62 2.59  
## 13 0.56 1.23 1.37 1.59 1.74 1.81 2.05 2.10 2.31 2.47 2.57 2.62 3.05 2.74 2.77  
## 14 0.56 1.19 1.32 1.54 1.70 1.76 2.00 2.06 2.28 2.45 2.57 2.62 2.74 3.10 2.83  
## 15 0.56 1.05 1.17 1.37 1.53 1.59 1.84 1.90 2.15 2.35 2.50 2.59 2.77 2.83 3.31  
##   
## [[2]]  
## 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15  
## 1 1.00 0.45 0.43 0.41 0.39 0.39 0.38 0.37 0.36 0.35 0.35 0.35 0.34 0.34 0.32  
## 2 0.45 1.00 0.81 0.81 0.80 0.80 0.77 0.76 0.71 0.67 0.63 0.60 0.54 0.51 0.44  
## 3 0.43 0.81 1.00 0.83 0.82 0.82 0.80 0.79 0.75 0.70 0.66 0.64 0.57 0.55 0.47  
## 4 0.41 0.81 0.83 1.00 0.84 0.84 0.83 0.82 0.79 0.75 0.72 0.69 0.63 0.60 0.52  
## 5 0.39 0.80 0.82 0.84 1.00 0.85 0.85 0.84 0.82 0.79 0.75 0.73 0.67 0.64 0.56  
## 6 0.39 0.80 0.82 0.84 0.85 1.00 0.85 0.85 0.83 0.80 0.77 0.75 0.68 0.66 0.58  
## 7 0.38 0.77 0.80 0.83 0.85 0.85 1.00 0.87 0.86 0.84 0.82 0.80 0.74 0.72 0.64  
## 8 0.37 0.76 0.79 0.82 0.84 0.85 0.87 1.00 0.86 0.85 0.83 0.81 0.76 0.74 0.66  
## 9 0.36 0.71 0.75 0.79 0.82 0.83 0.86 0.86 1.00 0.87 0.86 0.85 0.81 0.79 0.72  
## 10 0.35 0.67 0.70 0.75 0.79 0.80 0.84 0.85 0.87 1.00 0.88 0.87 0.85 0.83 0.78  
## 11 0.35 0.63 0.66 0.72 0.75 0.77 0.82 0.83 0.86 0.88 1.00 0.88 0.87 0.86 0.81  
## 12 0.35 0.60 0.64 0.69 0.73 0.75 0.80 0.81 0.85 0.87 0.88 1.00 0.88 0.87 0.83  
## 13 0.34 0.54 0.57 0.63 0.67 0.68 0.74 0.76 0.81 0.85 0.87 0.88 1.00 0.89 0.87  
## 14 0.34 0.51 0.55 0.60 0.64 0.66 0.72 0.74 0.79 0.83 0.86 0.87 0.89 1.00 0.88  
## 15 0.32 0.44 0.47 0.52 0.56 0.58 0.64 0.66 0.72 0.78 0.81 0.83 0.87 0.88 1.00

## Question2 survival analysis

* use the same dataset
* relationships between baseline and survival
* create an individual level dataset

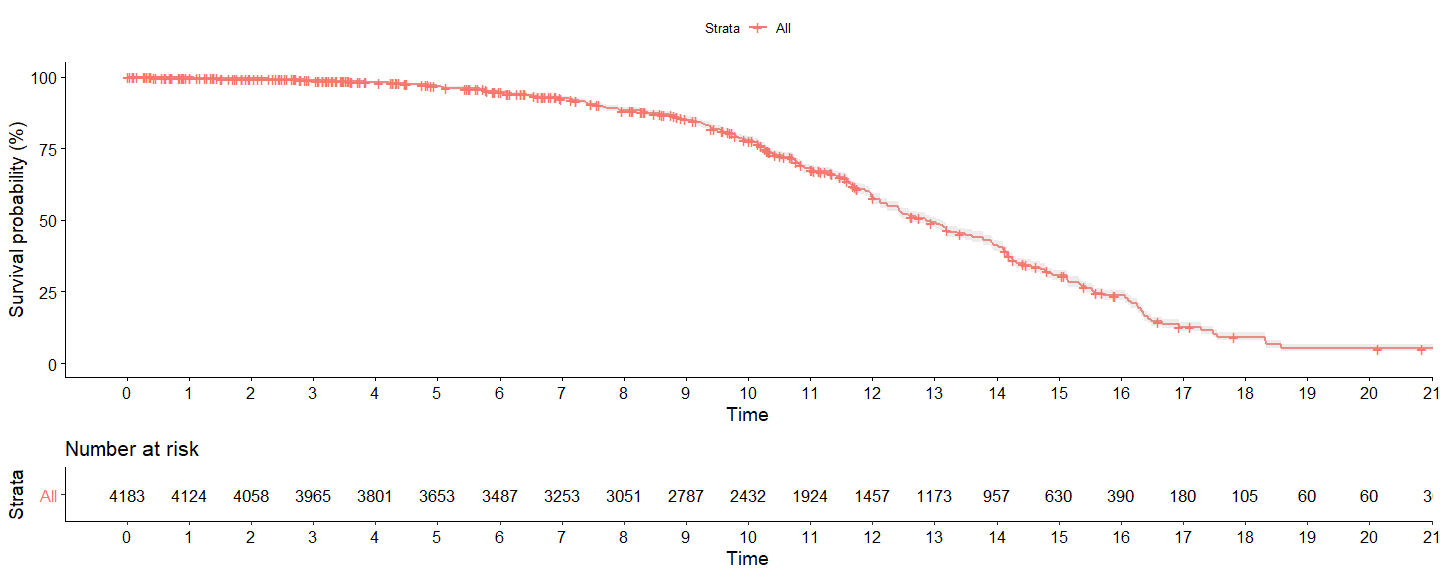
### a. Kaplan-Meier survival curves

* overall and operation type

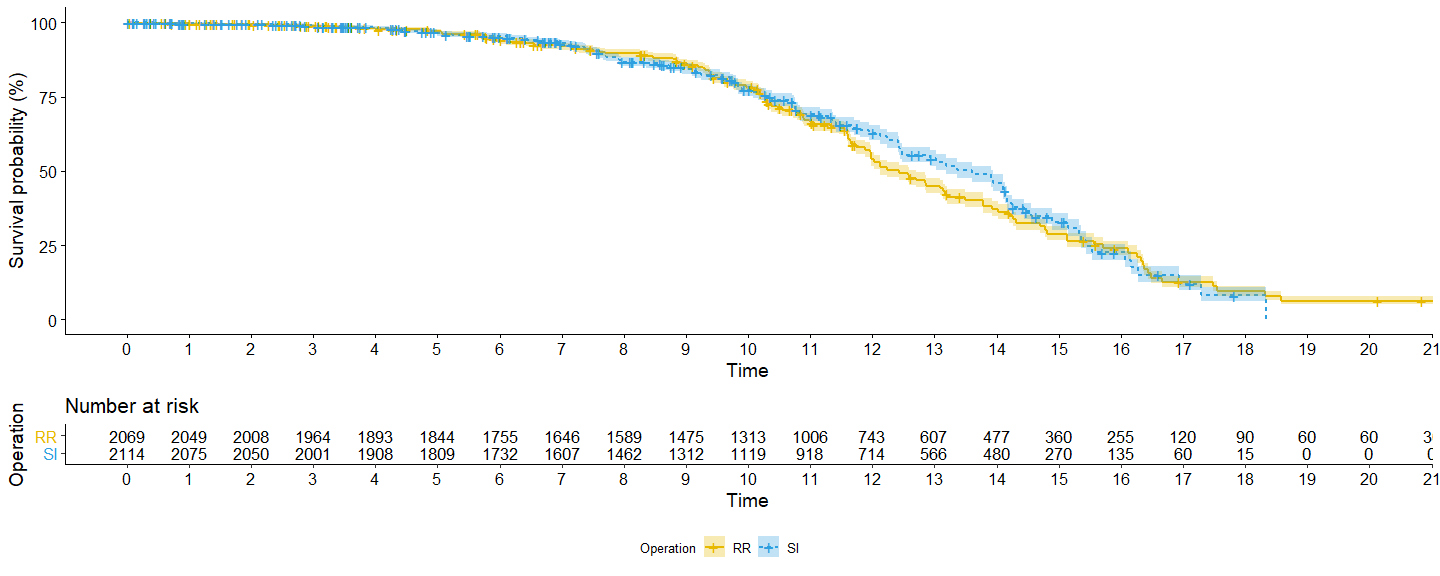
aort1$surobj <- with(aort1, Surv(ev\_time, event == 1))  
  
fit0 <- survfit(surobj ~ 1,  
 data = aort1,  
 type = "kaplan-meier")  
fit1 <- survfit(surobj ~ type\_op,  
 data = aort1,  
 type = "kaplan-meier")

plot\_fit0 <- fit0 %>%  
 survminer::ggsurvplot(  
 data = aort1,  
 break.time.by = 1,  
 conf.int = TRUE,  
 fun = "pct",  
 risk.table = TRUE,  
 size = 1,  
 linetype = "strata")  
  
plot\_fit1 <- fit1 %>%  
 survminer::ggsurvplot(  
 data = aort1,  
 break.time.by = 1,  
 conf.int = TRUE,  
 fun = "pct",  
 risk.table = TRUE,  
 size = 1,  
 linetype = "strata",  
 palette= c("#E7B800",  
 "#2E9FDF"),  
 legend = "bottom",  
 legend.title = "Operation",  
 legend.labs = c("RR", "SI"))

plot\_fit0



plot\_fit1



### b. result from last question

* estimate the predicted survival at 10 years for each type
* approximate the estimate of the hazard ratio

For operation RR, the estimate survival at 10 years is 0.784; for operation IS, the estimate survival at 10 years is 0.0.772. So based on the approximation for median survival time and the survival at 10 years.  
The proportional hazard ratio for RR and SI is around 1.08.

summary(fit1, times = 10)

## Call: survfit(formula = surobj ~ type\_op, data = aort1, type = "kaplan-meier")  
##   
## type\_op=RR   
## time n.risk n.event survival std.err lower 95% CI   
## 1.00e+01 1.31e+03 3.87e+02 7.84e-01 9.79e-03 7.65e-01   
## upper 95% CI   
## 8.03e-01   
##   
## type\_op=SI   
## time n.risk n.event survival std.err lower 95% CI   
## 1.00e+01 1.12e+03 3.84e+02 7.72e-01 1.03e-02 7.52e-01   
## upper 95% CI   
## 7.93e-01

fit1

## Call: survfit(formula = surobj ~ type\_op, data = aort1, type = "kaplan-meier")  
##   
## n events median 0.95LCL 0.95UCL  
## type\_op=RR 2069 1350 12.4 12.1 12.7  
## type\_op=SI 2114 1046 13.6 13.2 13.9

# survdiff(surobj ~ type\_op, data = aort1)

h\_rr <- (log(0.5) - log(0.784)) / (12.4 - 6.42)  
h\_si <- (log(0.5) - log(0.772)) / (13.6 - 7.39)  
h\_rr / h\_si

## [1] 1.075337

### c. categorize the baseline value into 4 group

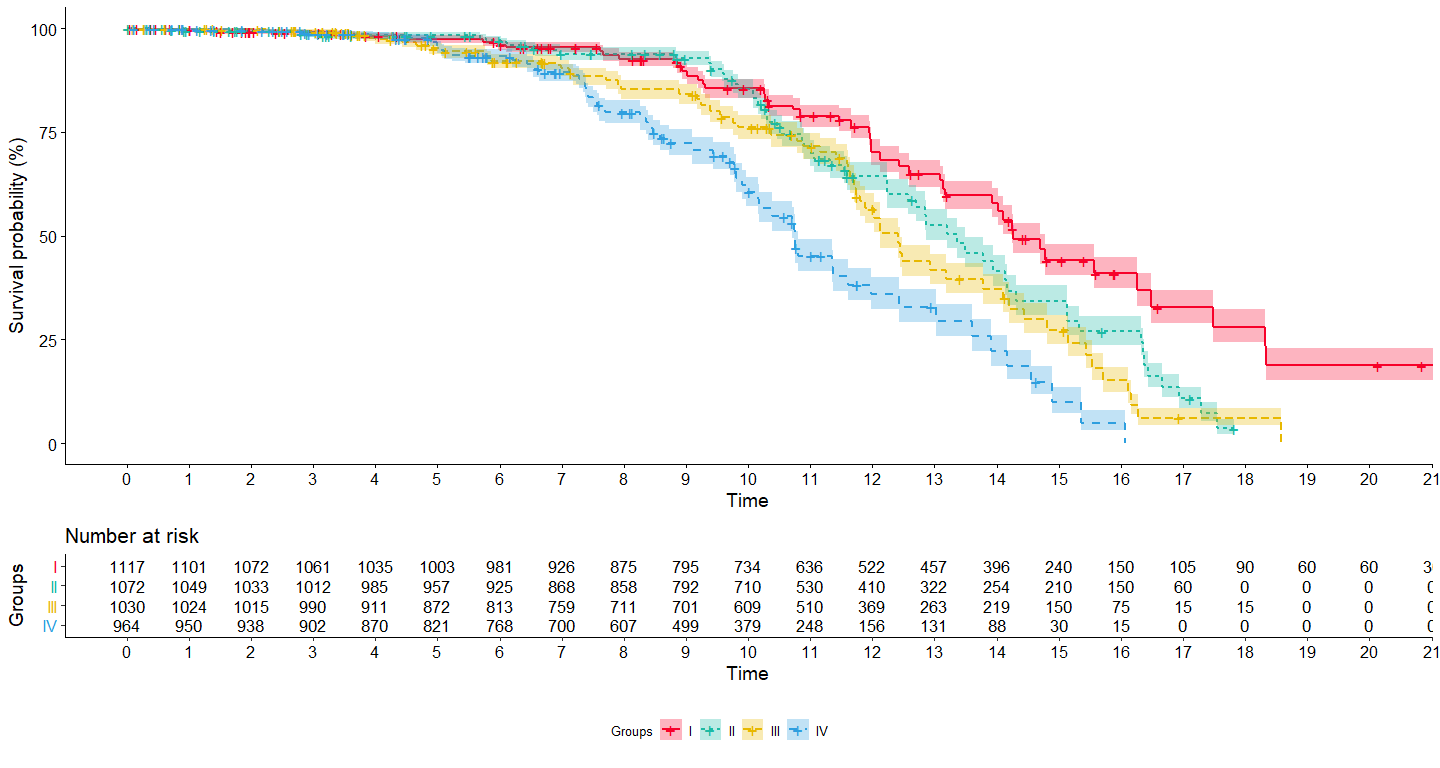
#' show\_table() to get a frequency table  
#'  
#' @param data the dataset  
#' @param arg the categorical variable interested  
#' @return NULL but print function  
#' @examples  
#' show\_table(aort1, sex)  
show\_table <- function(data, arg) {  
 table <- with(data, table(arg)) %>%  
 as.data.frame()   
 print(table)  
}  
  
aort3 <- aort %>%   
 mutate(baseline = ifelse(time == 0, ao\_gradient, NA)) %>%  
 fill(baseline) %>%  
 group\_by(id, baseline) %>%  
 nest() %>%  
 mutate(group = case\_when(  
 baseline <= quantile(.$baseline, 0.25) ~ "1",  
 baseline <= quantile(.$baseline, 0.50) ~ "2",  
 baseline <= quantile(.$baseline, 0.75) ~ "3",  
 baseline <= quantile(.$baseline, 1.00) ~ "4")) %T>%  
 show\_table(.$group) %>%  
 unnest()

## arg Freq  
## 1 1 125  
## 2 2 125  
## 3 3 125  
## 4 4 125

aort3$surobj <- with(aort3, Surv(ev\_time, event == 1))  
  
fit2 <- survfit(surobj ~ group,  
 data = aort3)  
  
plot\_fit2 <- fit2 %>%  
 survminer::ggsurvplot(  
 data = aort3,  
 break.time.by = 1,  
 conf.int = TRUE,  
 fun = "pct",  
 risk.table = TRUE,  
 size = 1,  
 linetype = "strata",  
 palette= c("#f7022a",   
 "#1cbaa4",  
 "#E7B800",  
 "#2E9FDF"),  
 legend = "bottom",  
 legend.title = "Groups",  
 legend.labs = c("I", "II", "III", "IV"))

According to the Kaplan-Meier plot (plot\_fit2), there is a clear correlation between the baseline aortic gradient level and the survival results. In population level, the lower baseline gradient the subject suffered, the lower risk rate and the longer of the median survival time.

plot\_fit2



fit2

## Call: survfit(formula = surobj ~ group, data = aort3)  
##   
## n events median 0.95LCL 0.95UCL  
## group=1 1117 554 14.3 14.2 14.8  
## group=2 1072 644 13.4 12.9 13.5  
## group=3 1030 636 12.4 12.1 12.4  
## group=4 964 562 10.8 10.7 10.8

### d. fit a cox model and only use operation

Based on the Cox proportional hazard model (cox1), the operation type does not have a significant effect on the patient’s survival time (p = 0.37). We cannot reject the null hypothesis. According to the model, the log risk rate will decrease 0.037, which is equivalent to 0.9635 (95% CI: 0.8875, 1.046) fold lower risk, for patient in SI operation than the RR operation.

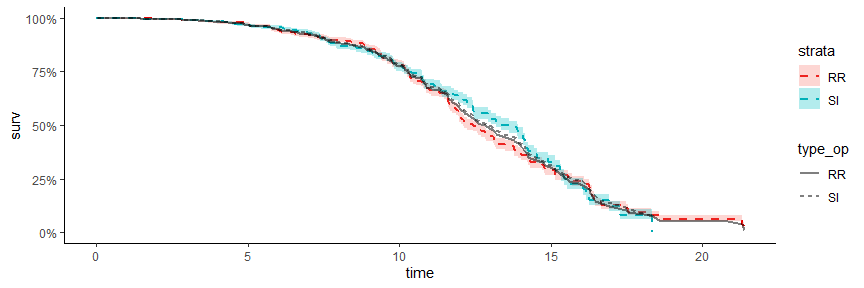
cox1 <- coxph(surobj ~ type\_op,   
 data = aort1)  
summary(cox1)

## Call:  
## coxph(formula = surobj ~ type\_op, data = aort1)  
##   
## n= 4183, number of events= 2396   
##   
## coef exp(coef) se(coef) z Pr(>|z|)  
## type\_opSI -0.03723 0.96346 0.04190 -0.888 0.374  
##   
## exp(coef) exp(-coef) lower .95 upper .95  
## type\_opSI 0.9635 1.038 0.8875 1.046  
##   
## Concordance= 0.511 (se = 0.006 )  
## Likelihood ratio test= 0.79 on 1 df, p=0.4  
## Wald test = 0.79 on 1 df, p=0.4  
## Score (logrank) test = 0.79 on 1 df, p=0.4

### e. predict the models cox1 and fit1

As seen in the overlay plot (plot\_cox), the colored dashed lines represent the Kaplan-Meier curves, and the grey solid lines represent the Cox predict curves. Due to no cross-over for different operation types and seemingly constant decreasing on survival rate. The Cox model is reasonable to be used for the model fitting. However due to the constant hazard ratio assumption for cox model, certain violations on the assumption for different operational type might happened. There is a shrinkage effect toward the mean of the overall survival trend.

aort4 <- aort1 %>%  
 mutate(coxsur = predict(cox1, type = "survival"))  
  
plot\_cox <- autoplot(fit1, data = aort1,  
 surv.size = 1,  
 surv.linetype = "dashed",  
 censor = FALSE) +  
 geom\_line(data = aort4,  
 aes(ev\_time, coxsur,   
 group = type\_op,  
 linetype = type\_op),  
 color = "black",  
 alpha = 0.50,  
 size = 1) +  
 scale\_colour\_hue(l = 45, c = 200) +  
 theme\_classic()   
  
plot\_cox



### f. add baseline value of sqrt\_aort\_grad

* add base\_sqrt\_aort into cox model

According to model cox2, both operation types and the baseline aortic gradient levels can significantly affect the subject survival time (pvalue << 0.001). In population level, on average, patients with the SI operation will experience 0.512 decreasing on log risk rate than RR operation; in another words, the risk rate will decrease to 0.600 (95% CI: 0.546, 0.658) fold in SI patients than RR patients. SI patients can enjoy a better survival time. Also, on average, every unit increase on baseline sqrt aortic gradient level will cause increase of risk rate to 1.733 (95% CI: 1.656, 1.814) folds. This indicates that the patients with higher baseline aortic gradient suffer higher risk of event.

aort5 <- aort3 %>%   
 mutate(basesqrt = sqrt(baseline))  
  
cox2 <- coxph(surobj ~ type\_op + basesqrt,   
 data = aort5)  
tidyc2 <- broom.mixed::tidy(cox2)

summary(cox2)

## Call:  
## coxph(formula = surobj ~ type\_op + basesqrt, data = aort5)  
##   
## n= 4183, number of events= 2396   
##   
## coef exp(coef) se(coef) z Pr(>|z|)   
## type\_opSI -0.51188 0.59937 0.04728 -10.83 <2e-16 \*\*\*  
## basesqrt 0.54982 1.73293 0.02322 23.68 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## exp(coef) exp(-coef) lower .95 upper .95  
## type\_opSI 0.5994 1.6684 0.5463 0.6576  
## basesqrt 1.7329 0.5771 1.6558 1.8136  
##   
## Concordance= 0.638 (se = 0.007 )  
## Likelihood ratio test= 577.3 on 2 df, p=<2e-16  
## Wald test = 560.6 on 2 df, p=<2e-16  
## Score (logrank) test = 559.5 on 2 df, p=<2e-16

tidyc2

## # A tibble: 2 x 5  
## term estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 type\_opSI -0.512 0.0473 -10.8 2.56e- 27  
## 2 basesqrt 0.550 0.0232 23.7 6.35e-124