weekly_july5

Dohun Lee

2022-07-29

Hello Solon, this is a Rmarkdown to discuss two of the things I've been working on:

1. Two group statistical testings
2. KM curves

1.0 P-values

First, append purity and subclonality estimations from liquidCNA. **Then, group subclonalities depending on Progression is YES or NO.**

```
RECIST <- append.RECIST()
yes.prog <- which(RECIST$Progression == "YES")
no.prog <- which(RECIST$Progression == "NO")

yes.progression <- RECIST[yes.prog,]$rat
no.progression <- RECIST[no.prog,]$rat

no.progression <- no.progression[!(is.na(no.progression))]
yes.progression <- yes.progression[!(is.na(yes.progression))]
```

T-test:

First, using shapiro to test for normality:

```
shapiro.test(no.progression) # p-value < 0.05
```

```
##
##  Shapiro-Wilk normality test
##
## data:  no.progression
## W = 0.75785, p-value = 1.234e-09
```

```
shapiro.test(yes.progression) # p-value < 0.05
```

```
##
##  Shapiro-Wilk normality test
##
## data:  yes.progression
## W = 0.44862, p-value < 2.2e-16
```

p-value is less than 0.05 implying that the distribution of the data are significantly different from normal distribution. Thus, the distribution of estimated subclonalities are skewed, **normality should not be assumed and T-test is inappropriate**.

Still, doing the t-test, p-value is 0.3489

```r
t.test(no.progression, yes.progression, var.equal = FALSE) #p-value > 0.05
```

```
##
##  Welch Two Sample t-test
##
## data:  no.progression and yes.progression
## t = -0.93905, df = 196.34, p-value = 0.3489
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3191258  0.1132459
## sample estimates:
## mean of x mean of y
## 0.4137283 0.5166682
```

As normality cannot be assumed, Unpaired Two-Samples **Wilcoxon Test** was conducted instead. **This however resulted a p-value of 1**

```r
wilcox.test(no.progression, yes.progression, alternative = "two.sided")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  no.progression and yes.progression
## W = 4671.5, p-value = 1
## alternative hypothesis: true location shift is not equal to 0
```

As opposed to this **purity estimation results significant for both** T-test and Wilcox test:

```r
t.test(RECIST[no.prog,]$purity_mean, RECIST[yes.prog,]$purity_mean, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  RECIST[no.prog, ]$purity_mean and RECIST[yes.prog, ]$purity_mean
## t = -2.8814, df = 160.57, p-value = 0.004501
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.0951725 -0.0177658
## sample estimates:
## mean of x mean of y
## 0.1143836 0.1708527
```

```r
wilcox.test(RECIST[no.prog,]$purity_mean, RECIST[yes.prog,]$purity_mean, alternative = "two.sided")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  RECIST[no.prog, ]$purity_mean and RECIST[yes.prog, ]$purity_mean
## W = 3349, p-value = 0.0004655
## alternative hypothesis: true location shift is not equal to 0
```

With Wilcox testing giving a p-value of 1, I am wondering if I have made an error in doing the statistic test somewhere. Do you have any idea what may be the cause here?

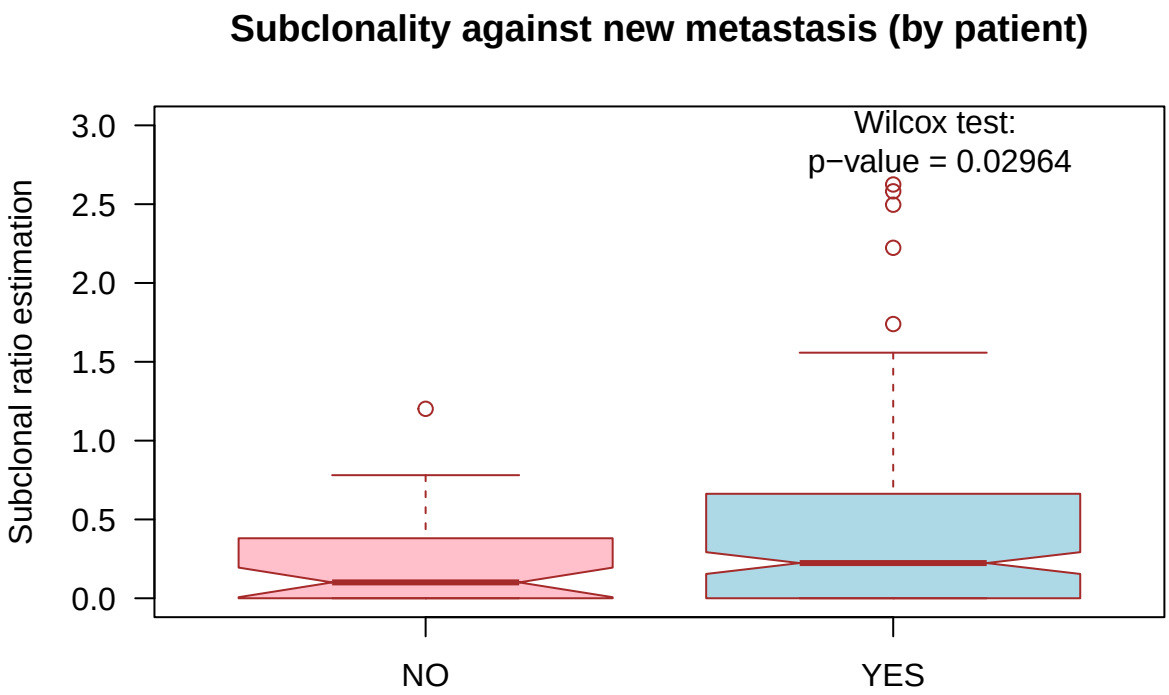
2.0 Using subclonality to estimate metastasis

For each patient, they were grouped into whether they have any new metastasis ('y' in RECIST) or not. Then, for each group, their respective subclonality estimations were grouped as well:

```
#function to categorise patient as newly metastasised or not
patient.metastasis <- function(p){
  p.id <- patient_ids[p]
  p.RECIST <- RECIST[which(RECIST$Patient_ID == p.id),]
  m <- "y" %in% unname(unlist(p.RECIST[,5:14]))
  return(m)
}

#use the function and group...
p.metastasis <- sapply(1:80, function(x) patient.metastasis(x))

#group subclonality estimations respectively....
subclonalities <- (sapply(1:length(liquidCNA_results),
                    function(x) as.numeric(liquidCNA_results[[x]]$rat)))
yes.met.rat <- unlist(sapply(which(p.metastasis), function(x) subclonalities[[x]]))
no.met.rat <- unlist(sapply(which(p.metastasis==F), function(x) subclonalities[[x]]))
```

Doing the Wilcox test, **we get a significant result**. I.e., distribution of subclonalities between patient groups of further metastasis and not, significantly differs:

```
p.met.res <- wilcox.test(yes.met.rat, no.met.rat, alternative = "two.sided")

boxplot(no.met.rat,
        yes.met.rat,
        ylim = c(0, 3),
        main = "Subclonality against new metastasis (by patient)",
        at = c(1,2),
        names = c("NO", "YES"),
        las = 1,
        col = c("pink","lightblue"),
        border = "brown",
        horizontal = FALSE,
        notch = TRUE,
        ylab = "Subclonal ratio estimation")
text(2.1, 2.9, paste0("Wilcox test: \np-value = ",round(p.met.res$p.value, 5)))
```



Thus, survival analysis was done next

3.0 Survival analysis

load Data and package

```
library("survival")
library("survminer")
```

```
##
## Attaching package: 'survminer'
```

```
## The following object is masked from 'package:survival':
##
##     myeloma
```

```
load("../DATA/survival_df.RData")
```

Survival.df

```
head(survival.df2)
```

```
##    id tstart tstop status event sub.group  subclone purity
## 1 339      0    15      1     2         1 0.2101278  0.050
## 2 339     15   256      0     1         1 0.2627656  0.050
## 3 388      0    17      1     2         2 0.5097418  0.130
## 4 388     17   203      1     1         1 0.0000000  0.125
## 5 388    203   736      1     1         2 0.9205857  0.065
## 6 388    736  1009      0     1         2 0.8145709  0.050
```

1. The first column are patient ids
2. tstart, tstop are in days of event. **Days start from RECIST's "Date.Meta" column. Could you confirm if this is the date of first primary metastasis for the patient please?**
3. status: 1 is recurrent event; 0 is right censored data
4. Event: 2 is metastasis; 1 is no event

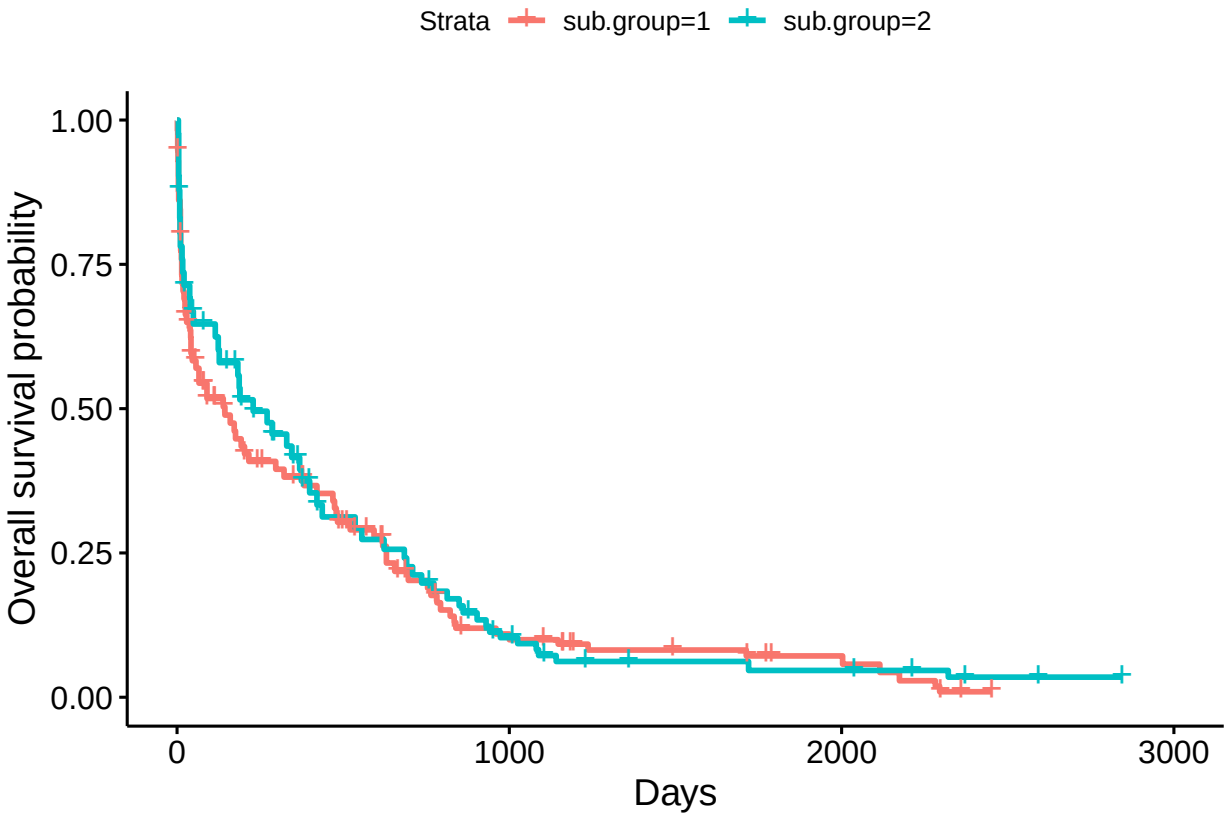
Furthermore, there are three covariates included in the data frame: sub.group (whether level of subclonality is high (2) or low (1)), subclonality ratio estimate and purity estimate. For sub.group, 2 is high subclonality group; and 1 is low subclonality group. Group2 has subclonality > 0.3656956 and 1 below it.

KM was done first:

```
model.2 = survfit(Surv(tstart,tstop,status) ~ sub.group,  data = survival.df2)
```

```
## Warning in Surv(tstart, tstop, status): Stop time must be > start time, NA
## created
```

```
ggsurvplot(model.2, data = survival.df2,
     xlab = "Days",
     ylab = "Overall survival probability")
```



The problem with KM is that events are single-off. So even if metastasis (our event in question) reoccurs, the analysis doesnot take this into account when computing. Therefore, Cox proportional hazards regression model (which allows for recurrent events) was explored.

The three covariates were used to model the Cox regression. Again, **purity was the only factor coming out to be significant**

```
model.1 = coxph(Surv(tstart,tstop,status) ~ (sub.group + subclone + purity + cluster(id)),
                method="breslow",
                robust=TRUE,
                data = survival.df2)
```

```
## Warning in Surv(tstart, tstop, status): Stop time must be > start time, NA
## created
```

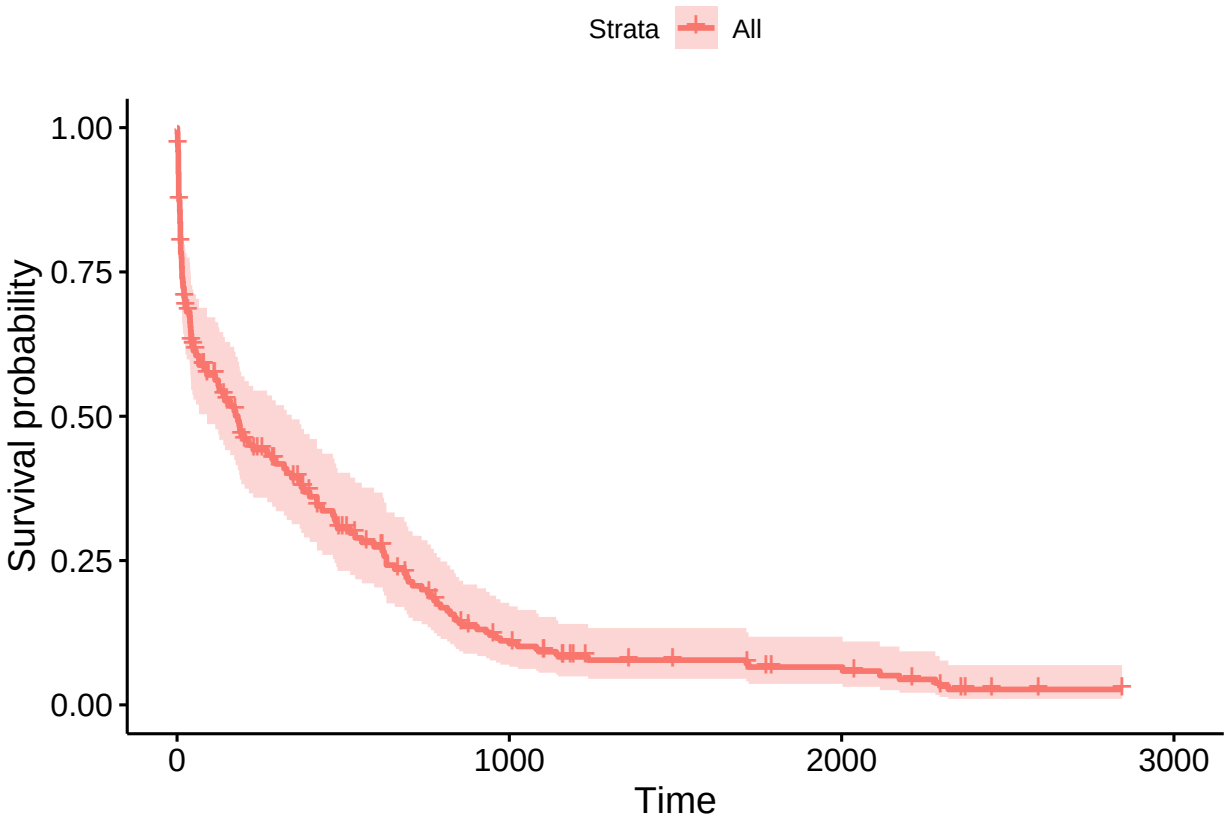
```
summary(model.1)
```

```
## Call:
## coxph(formula = Surv(tstart, tstop, status) ~ sub.group + subclone +
##     purity, data = survival.df2, robust = TRUE, method = "breslow",
##     cluster = id)
##
##   n= 194, number of events= 125
##    (27 observations deleted due to missingness)
##
##              coef exp(coef) se(coef) robust se      z Pr(>|z|)
## sub.group  0.3078    1.3604   0.2823    0.3020  1.019   0.3081
## subclone  -0.2639    0.7681   0.3166    0.2886 -0.914   0.3606
## purity    -1.6001    0.2019   0.7419    0.6674 -2.398   0.0165 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## sub.group    1.3604     0.7351   0.75267    2.4588
## subclone     0.7681     1.3020   0.43626    1.3523
## purity       0.2019     4.9534   0.05458    0.7467
##
## Concordance= 0.555  (se = 0.031 )
## Likelihood ratio test= 6.76  on 3 df,   p=0.08
## Wald test            = 6.93  on 3 df,   p=0.07
## Score (logrank) test = 6.1  on 3 df,   p=0.1,   Robust = 4.84  p=0.2
##
##   (Note: the likelihood ratio and score tests assume independence of
##      observations within a cluster, the Wald and robust score tests do not).
```

Plotting the survival(metastasis) probability...

```
ggsurvplot(survfit(model.1), data = survival.df2)
```

```
## Warning in Surv(tstart, tstop, status): Stop time must be > start time, NA
## created
```

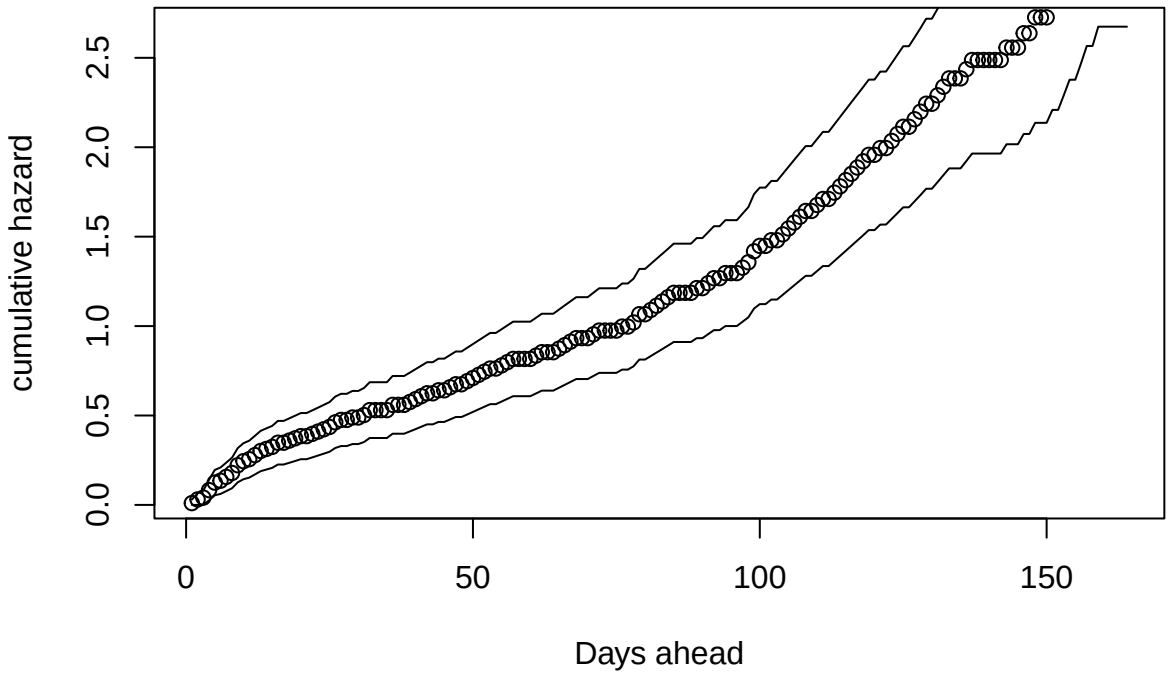


Plotting the cumulative hazard rate...

```
sfit <- survfit(model.1)
```

```
## Warning in Surv(tstart, tstop, status): Stop time must be > start time, NA
## created
```

```
cumhaz.upper <- -log(sfit$upper)
cumhaz.lower <- -log(sfit$lower)
cumhaz <- sfit$cumhaz # same as -log(sfit$surv)

plot(cumhaz, xlab="Days ahead", ylab="cumulative hazard",
     ylim=c(min(cumhaz.lower), max(cumhaz.upper)))
lines(cumhaz.lower)
lines(cumhaz.upper)
```


cumulative hazard
0.0
0.5
1.0
1.5
2.0
2.5
0
50
100
150
Days ahead