Project 1

Juliette Decugis, Vinay Gautam, Adam Mills, Chenxi Yao

2022-09-26

Project 1: Discover the genes associated with survival time in a common cancer

Loading and Renaming the Data

```
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(data.table)
```

```
##
## Attaching package: 'data.table'
##
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
##
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
library(skimr)
library(dplyr)
library(vdocs)
library(grid)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
DATA_DIR <- "~/Downloads/HW1"

mrna_orig <- fread(file.path(DATA_DIR, "data_mrna_agilent_microarray.txt")) %>%
  as_tibble()
clinical_orig <- fread(file.path(DATA_DIR, "brca_metabric_clinical_data.tsv")) %>%
  as_tibble()
```

```
colnames(clinical_orig) <- c("Study_ID","Patient_ID","Sample_ID","Age_at_Diagnosis",
                             "Type_of_Breast_Surgery","Cancer_Type","Cancer_Type_Detailed","Cellularity
                             "Chemotherapy","Pam50PlusClaudinMinusLow_Subtype","Cohort",
                             "ER_status_measured_by_IHC","ER_Status","Neoplasm_Histologic_Grade",
                             "HER2_status_measured_by_SNP6","HER2_Status","Tumor_Other_Histologic_Subty
                             "Hormone_Therapy","Inferred_Menopausal_State","Integrative_Cluster",
                             "Primary_Tumor_Laterality","Lymph_nodes_examined_positive","Mutation_Count
                             "Nottingham_prognostic_index","Oncotree_Code","Overall_Survival_Months",
                             "Overall_Survival_Status","PR_Status","Radio_Therapy","Relapse_Free_Status_
                             "Relapse_Free_Status","Number_of_Samples_Per_Patient","Sample_Type","Sex",
                             "3Gene_classifier_subtype","TMB_nonsynonymous","Tumor_Size","Tumor_Stage",
                             "Patient_Vital_Status")

#Study_ID only has 1 value, so we can drop it.
#Patient_ID is always equal to Sample_ID, so can drop one of those too.
#Sex is all Female, Sample_Type is all the same, Number_of_Samples_Per_Patient,
clinical_clean_v1 <- clinical_orig %>% select(-Study_ID,-Sample_ID,-Sex,-Sample_Type,-Number_of_Samples_

#if you want to check frequency of elements in a table
#   The useNA part will make sure to include NA
# clinical_orig$Relapse_Free_Status %>% table(useNA = "always")


#Relapse_Free_Status into binary, Overall_Survival_Status into binary.
clinical_clean_v1 <- mutate(clinical_clean_v1,Relapse =
                              case_when(
                                Relapse_Free_Status == "0:Not Recurred" ~ 0,
                                Relapse_Free_Status == "1:Recurred" ~ 1
                              ), Died =
                              case_when(
                                Overall_Survival_Status == "0:LIVING" ~ 0,
                                Overall_Survival_Status == "1:DECEASED" ~ 1
                              )
)

#adding a column for the age when people die.
clinical_clean_v1 <- clinical_clean_v1 %>%
  mutate(age_death = case_when(
    Overall_Survival_Status == 1 ~ Overall_Survival_Months/12 + Age_at_Diagnosis
  )
```

```
  )


#a table for the general population life expectancy.
#At a certain age, this is how many years we expect a `normal` American
#woman to live.
# Taken from https://www.ssa.gov/oact/STATS/table4c6.html
life_exp_table <- tibble(age = c(20:97),
                         exp_years_left = c(
                           61.93,60.95,59.98,59.01,58.04,57.07,56.11,55.14,54.17,53.21,
                           52.25,51.29,50.34,49.38,48.43,47.48,46.53,45.59,44.64,43.7,42.76,41.82,
                           40.88,39.95,39.01,38.08,37.16,36.24,35.32,34.41,33.5,32.6,31.71,30.82,
                           29.93,29.06,28.19,27.33,26.48,25.63,24.79,23.96,23.14,22.32,21.51,20.7,
                           19.89,19.1,18.31,17.52,16.75,16,15.25,14.52,13.8,13.1,12.41,11.74,
                           11.08,10.45,9.83,9.23,8.65,8.09,7.56,7.05,6.56,6.1,5.67,5.26,4.88,
                           4.52,4.2,3.9,3.63,3.39,3.17,2.98)
)
#exp_life is the age we predict given their current age.
life_exp_table <- life_exp_table %>% mutate(total_exp_life = age + exp_years_left)


#One Attempt to normalize data. take the age_death and compare it to exp_life
#We cant use the people who haven't died yet for this metric. still have 1144 ppl
# that died.

#This makes a table of exp_years_left for each person
temp <- tibble(age=round(clinical_clean_v1$Age_at_Diagnosis))
temp <- left_join(temp,life_exp_table)
```

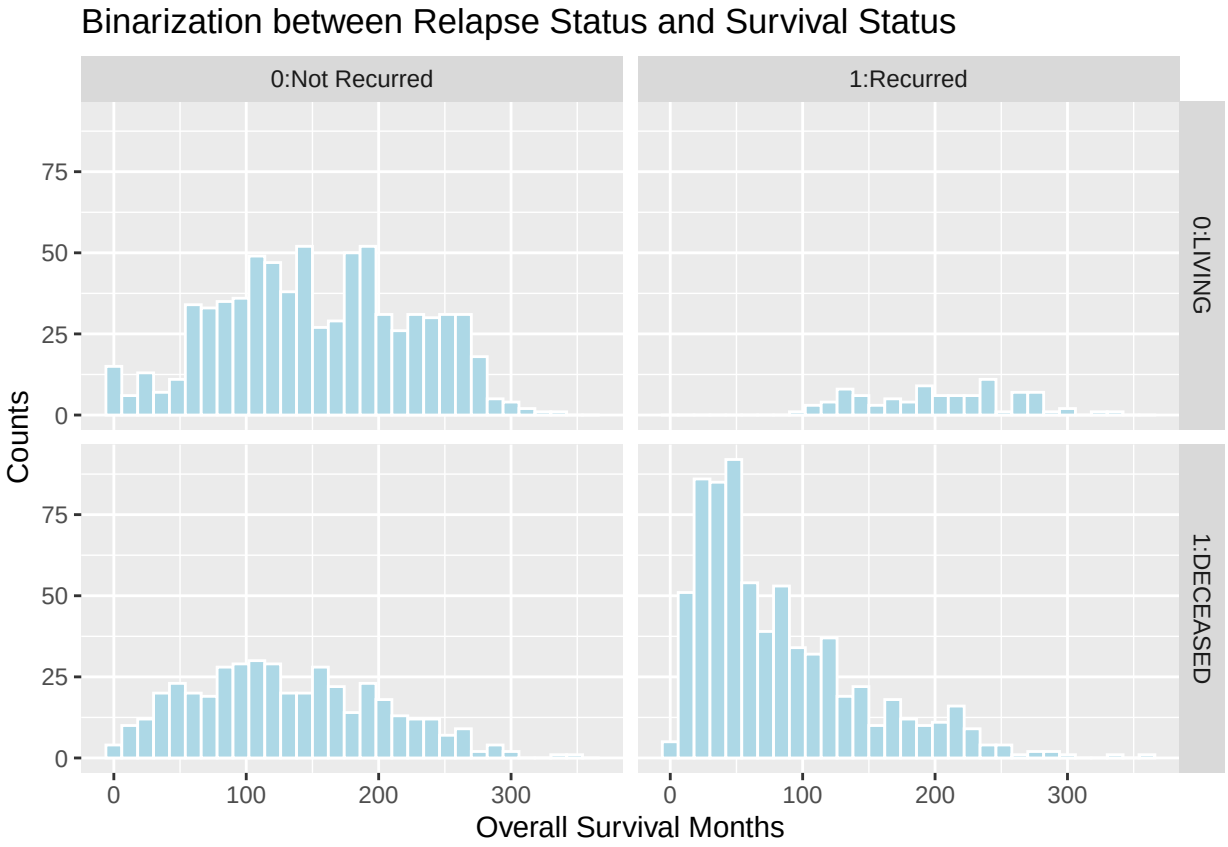
```
## Joining, by = "age"
```

```
#Add a column for diff in actual age of death vs life expectancy
clinical_clean_v1 <- clinical_clean_v1 %>% mutate(
  norm_life = (age_death - temp$total_exp_life )/ temp$total_exp_life
)
```

Filtering Data & EDA

For binarization, we take account in two factors, overall survival status and relapse status. We define the group of people who have no relapse and are still alive as "high survival"(Upper left in the graph) and the group of people who have relapse and are already dead as "low survival"(Lower right).

```
ggplot(data = filter(clinical_clean_v1, Relapse != 'NA', Died != 'NA'))+
  geom_histogram(aes(x = Overall_Survival_Months), fill = 'light blue', color = "white", binwidth = 12)+
  facet_grid(`Overall_Survival_Status`~`Relapse_Free_Status`)+
  labs(x = "Overall Survival Months",
       title = "Binarization between Relapse Status and Survival Status",
       y = "Counts")
```



Dropping NA survival months

As survival months has a very high influence on our low vs. high binarization, we decided it would be too risky to attempt to solve for the NA values.

```
sum(is.na(clinical_clean_v1$Overall_Survival_Months))
```

```
## [1] 528
```

```
clinical_clean_v2 <- clinical_clean_v1 %>% drop_na(Overall_Survival_Months)
clinical_clean_v2 %>% nrow()
```

```
## [1] 1981
```

We drop 528 patients with NA survival months.

Dropping Death due to Other Causes

We remove the patients who died of other causes, since those patients would be considered very low survival but don't reflect actual cancel progression.

```
clinical_clean_v2 <- clinical_clean_v2 %>% filter(!(Patient_Vital_Status == "Died of Other Causes"))
clinical_clean_v2 %>% nrow()
```

```
## [1] 1483
```

Dropping Recent Diagnosis

We remove patients who have survived less than 3 years but are still alive. We assumed those patients got recently diagnosed and may not properly reflect survival times.

```
clinical_clean_v2 <- clinical_clean_v2 %>% filter(!(Overall_Survival_Months <= 36 & Died == 0))
clinical_clean_v2 %>% nrow()
```

```
## [1] 1443
```

We are left with 1443 patients in our classification task.

About Relapse

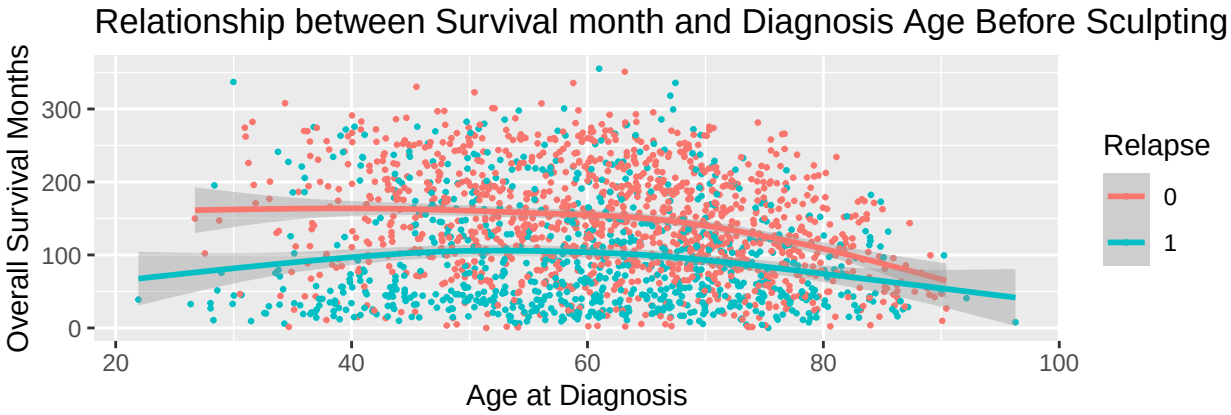
```
# Plot before sculpting
clinical_clean_v1$Relapse = as.character(clinical_clean_v1$Relapse)
plot_before <- ggplot(data = filter(clinical_clean_v1, Relapse != "NA")) +
  aes(y = `Overall_Survival_Months`, x = `Age_at_Diagnosis`,
      color = `Relapse`) +
  geom_point(size = 0.5) +
  geom_smooth()+
  labs(x = "Age at Diagnosis", y = "Overall Survival Months",
       title = "Relationship between Survival month and Diagnosis Age Before Sculpting")

# Plot after sculpting
clinical_clean_v2$Relapse = as.character(clinical_clean_v2$Relapse)
plot_after <- ggplot(data = filter(clinical_clean_v2, Relapse != "NA")) +
  aes(y = `Overall_Survival_Months`, x = `Age_at_Diagnosis`,
      color = `Relapse`) +
  geom_point(size = 0.5) +
  geom_smooth()+
  labs(x = "Age at Diagnosis", y = "Overall Survival Months",
       title = "Relationship between Survival month and Diagnosis After Sculpting")

grid.arrange(plot_before, plot_after, ncol = 1)
```

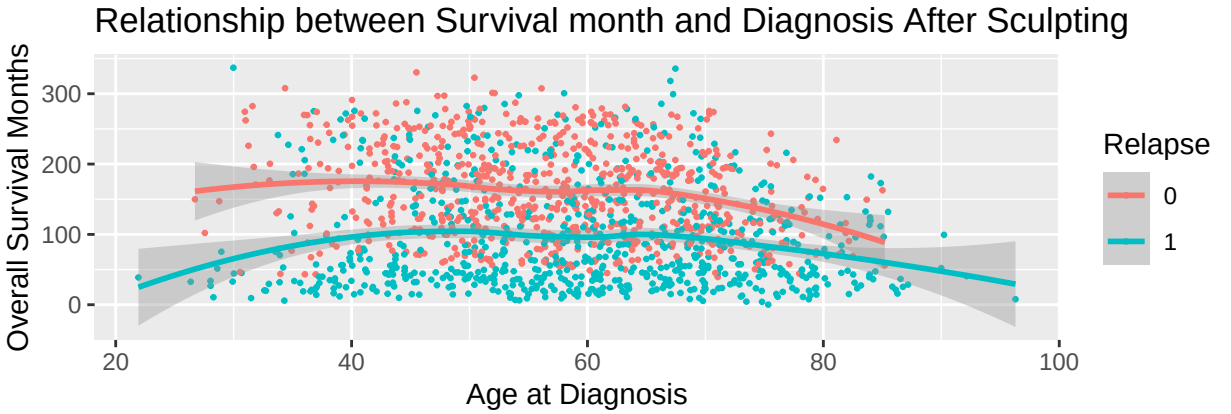
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 508 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 508 rows containing missing values (geom_point).
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```


Relationship between Survival month and Diagnosis Age Before Sculpting
Overall Survival Months
300
200
100
0
20
40
60
80
100
Age at Diagnosis
Relapse
0
1


Relationship between Survival month and Diagnosis After Sculpting
Overall Survival Months
300
200
100
0
20
40
60
80
100
Age at Diagnosis
Relapse
0
1