Breast cancer survival prediction

Yangyang Chen

2023-12-09

Description This is a dataset of breast cancer patients from a prospective study. Information including variables 1-14 were collected at the baseline, the column `Survival Months` records the length of following-up, and the column `Status` records the survival status of the patients at the end of their following-up. We are primarily interested in predicting the risk of death based on features 1-14. 1. Age

2. Race
3. Marital Status
4. T Stage: Adjusted AJCC 6th T
5. N Stage: Adjusted AJCC 6th N
6. 6th Stage: Breast Adjusted AJCC 6th Stage
7. Differentiate
8. Grade
9. A Stage: Regional — A neoplasm that has extended; Distant — A neoplasm that has spread to parts of the body remote from
10. Tumor Size: Each indicates exact size in millimeters.
11. Estrogen Status
12. Progesterone Status
13. Regional Node Examined
14. Reginol Node Positive
15. Survival Months
16. Status: Dead / Alive

Analytical goal (you may only address some of them but properly addressing more will improve the rate of your report):

1. Using variables 1-14 as the covariates to predict the risk of death.
2. Which factors (features) affect the risk significantly? Are there interacting effects?
3. Evaluate the performance of your model(s). Is your model achieving similar performance between the majority race group "White" and the minority "Black" (or "Black" + "Other")? If not, could you try to improve the fairness (i.e., reducing the gap of prediction performance between the majority and minority) of your model(s)?

Suggestions and tips: In the report, you should describe your final model and interpret its parameters in an accurate and useful manner. It is expected that you would first examine the *marginal distributions* and *pairwise relationships* between variables (e.g., to check to see whether any *nonlinearities* are immediately obvious), that you would explore several candidate models, and explain why you selected your model.

Also, you should check for violations of regression model assumptions, influential observations, multicollinearity, etc. This project may involve *logistic or survival models* not introduced in our class, which gives you a 5 points bonus. It would be great if you could be an active learner and figure out these challenges. In addition, model evaluation and fairness mentioned in the above Point 3 are interesting topics, on which you could try to explore. It would be helpful to be clear about your motivation for carrying out certain analyses as well as to be clear about interpretations of fitted model parameters. Your report should include a *table* summarizing parameter estimates associated with your final fitted model, characterizing predictor variables in a way that a reader can clearly understand.

Below you'll find some aspects to be addressed in your report. These are just a few suggestions, but feel free to add your own input/creativity to the analysis:

- Data exploration: descriptive statistics and visualization. You might want to, for instance:

o Include *a descriptive table* with summary statistics for all variables;

o Explore the *distribution of the outcome* and consider potential *transformations* (if necessary);

o See if there are any unusual observations and consider them as potential *outliers/influential points*.

- In your regression model, be watchful for variables that are highly *correlated* and be selective in the variables you will include in your analysis.
- Consider selective *interactions* between variables.
- DO NOT IGNORE *MODEL DIAGNOSTICS*

EDA

Methods

1. Data Description

```
cancer_df =
  read_csv("Project_2_data.csv") |>
  janitor::clean_names() |>
  as.data.frame()
```

```
## Rows: 4024 Columns: 16
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (11): Race, Marital Status, T Stage, N Stage, 6th Stage, differentiate, ...
## dbl  (5): Age, Tumor Size, Regional Node Examined, Reginol Node Positive, Su...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Change the factor variables
col_names =
  cancer_df |> select_if(is.character) |> colnames() |> as.vector()

cancer_df[,col_names] =
  lapply(cancer_df[,col_names], factor)

# Change character variable "Status" to binary variables
cancer_df =
  cancer_df |>
  mutate(
    status =
```

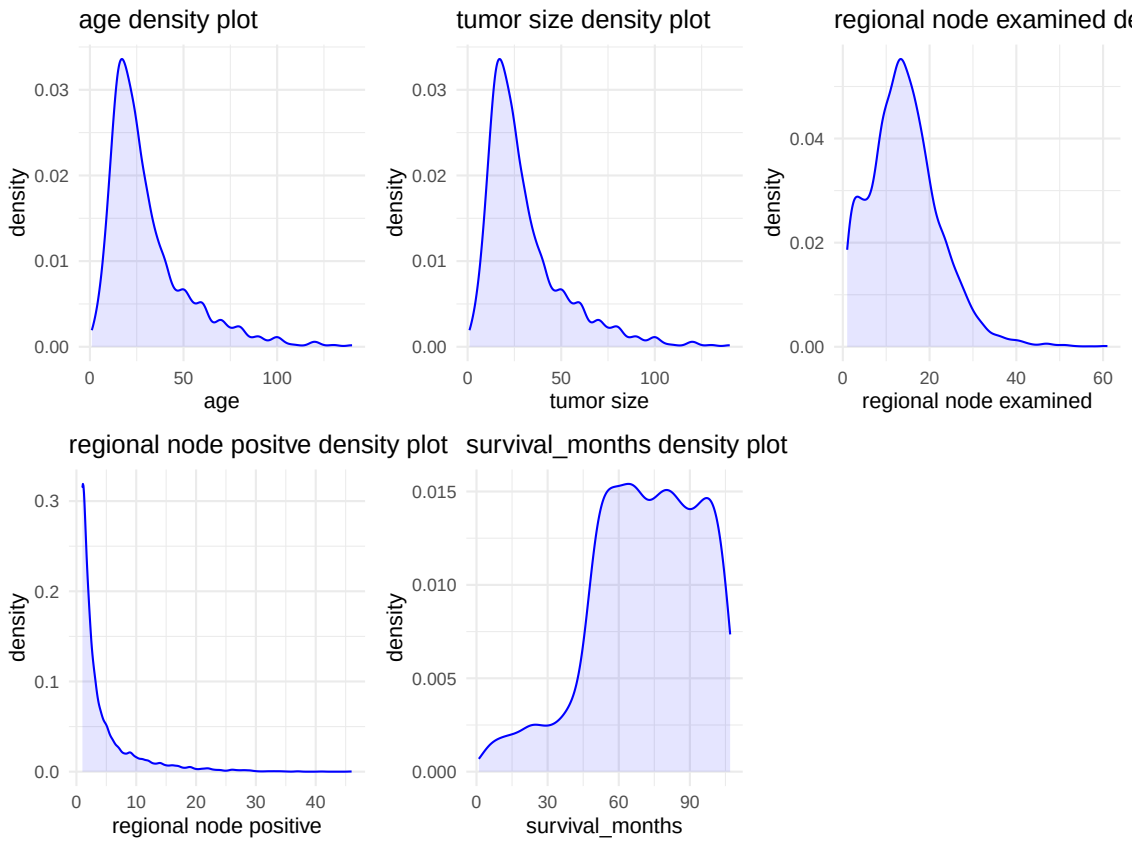
```
      case_match(
        status,
        "Alive" ~ 0,
        "Dead" ~ 1)
  )
```

The numeric descriptive statistics are:

age	tumor_size	regional_node_examined	reginol_node_positive	survival_months	status
Min. :30.00	Min. : 1.00	Min. : 1.00	Min. : 1.000	Min. : 1.0	Min. :0.0000
1st Qu.:47.00	1st Qu.: 16.00	1st Qu.: 9.00	1st Qu.: 1.000	1st Qu.: 56.0	1st Qu.:0.0000
Median :54.00	Median : 25.00	Median :14.00	Median : 2.000	Median : 73.0	Median :0.0000
Mean :53.97	Mean : 30.47	Mean :14.36	Mean : 4.158	Mean : 71.3	Mean :0.1531
3rd Qu.:61.00	3rd Qu.: 38.00	3rd Qu.:19.00	3rd Qu.: 5.000	3rd Qu.: 90.0	3rd Qu.:0.0000
Max. :69.00	Max. :140.00	Max. :61.00	Max. :46.000	Max. :107.0	Max. :1.0000

```
age_plt =
  cancer_df |>
  ggplot(aes(x = tumor_size)) +
  geom_density(color = "blue", fill = "blue", alpha = 0.1) +
  labs(
    title = "age density plot",
    x = "age"
  )

tumor_size_plt =
  cancer_df |>
  ggplot(aes(x = tumor_size)) +
  geom_density(color = "blue", fill = "blue", alpha = 0.1) +
  labs(
    title = "tumor size density plot",
    x = "tumor size"
  )

reg_node_exam_plt =
  cancer_df |>
  ggplot(aes(x = regional_node_examined)) +
  geom_density(color = "blue", fill = "blue", alpha = 0.1) +
  labs(
    title = "regional node examined density plot",
    x = "regional node examined"
  )

reg_node_pos_plt =
  cancer_df |>
  ggplot(aes(x = reginol_node_positive)) +
  geom_density(color = "blue", fill = "blue", alpha = 0.1) +
  labs(
```

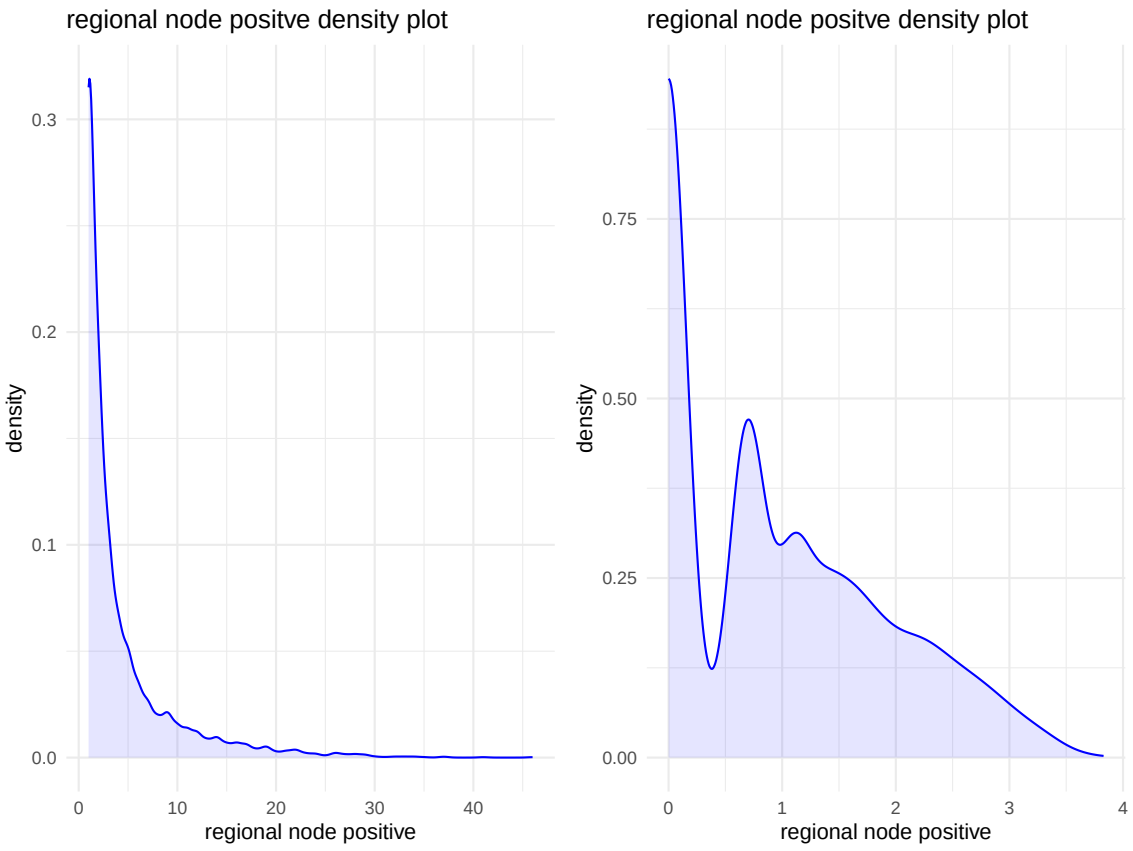
```
    title = "regional node positve density plot",
    x = "regional node positive"
  )

months_plt =
  cancer_df |>
  ggplot(aes(x = survival_months)) +
  geom_density(color = "blue", fill = "blue", alpha = 0.1) +
  labs(
    title = "survival_months density plot",
    x = "survival_months"
  )

gridExtra::grid.arrange(
  age_plt,
  tumor_size_plt,
  reg_node_exam_plt,
  reg_node_pos_plt,
  months_plt,
  nrow = 2,
  ncol = 3)
```



Comparing all numeric variables distribution, we have decided to conduct log-transform on `reginol_node_positive` variable:

```
org_reg_plt =
  cancer_df |>
  ggplot(aes(x = reginol_node_positive)) +
```

```
  geom_density(color = "blue", fill = "blue", alpha = 0.1) +
  labs(
    title = "regional node positve density plot",
    x = "regional node positive"
  )

log_reg_plt =
  cancer_df |>
  ggplot(aes(x = log(reginol_node_positive))) +
  geom_density(color = "blue", fill = "blue", alpha = 0.1) +
  labs(
    title = "regional node positve density plot",
    x = "regional node positive"
  )
gridExtra::grid.arrange(org_reg_plt, log_reg_plt, nrow = 1)
```



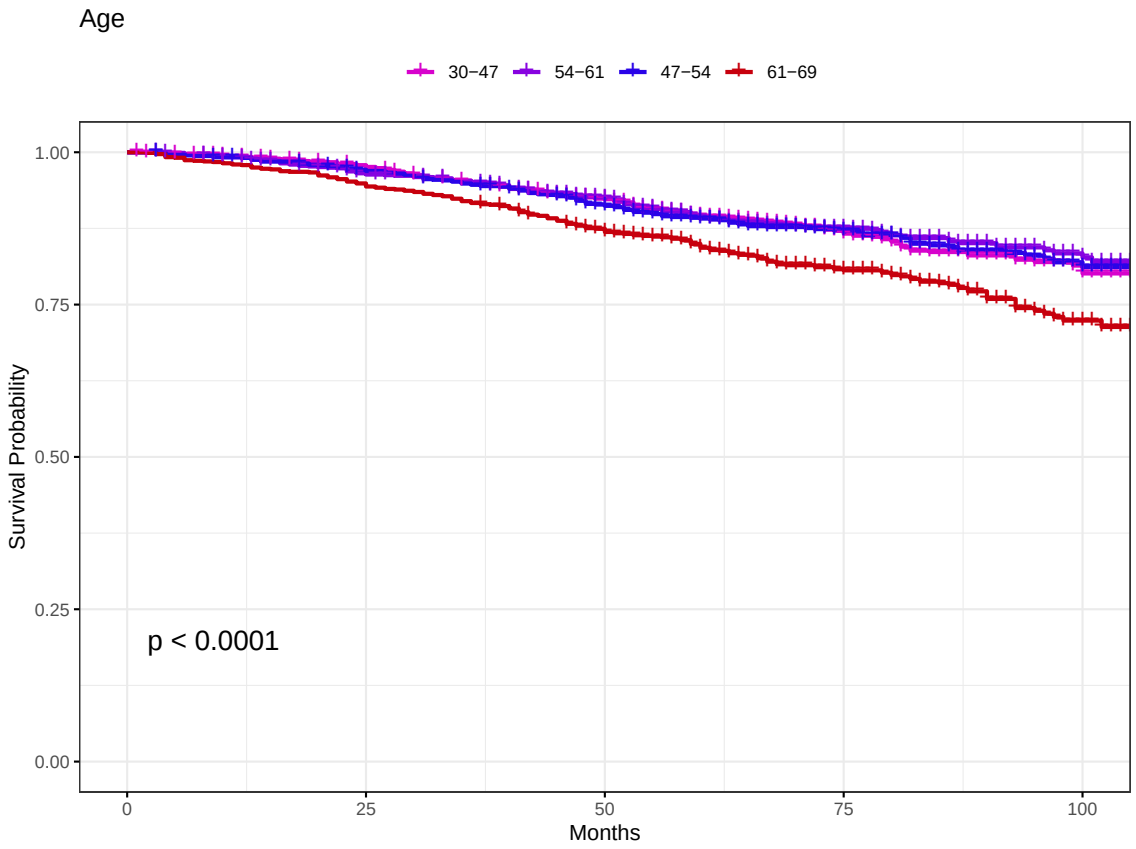
Comparing plots before and after log-transformation, we decided to abandoned log-transform on `reginol_node_positive` variable and adopt other methods to analyze.

Results

Univariate Analysis

Kaplan-Meier plots of each variable with a significant log rank test result. Non-significant Kaplan-Meier plots can be found in Appendix A.

```
KMplots_age
```




```
KMplots_tumor
```

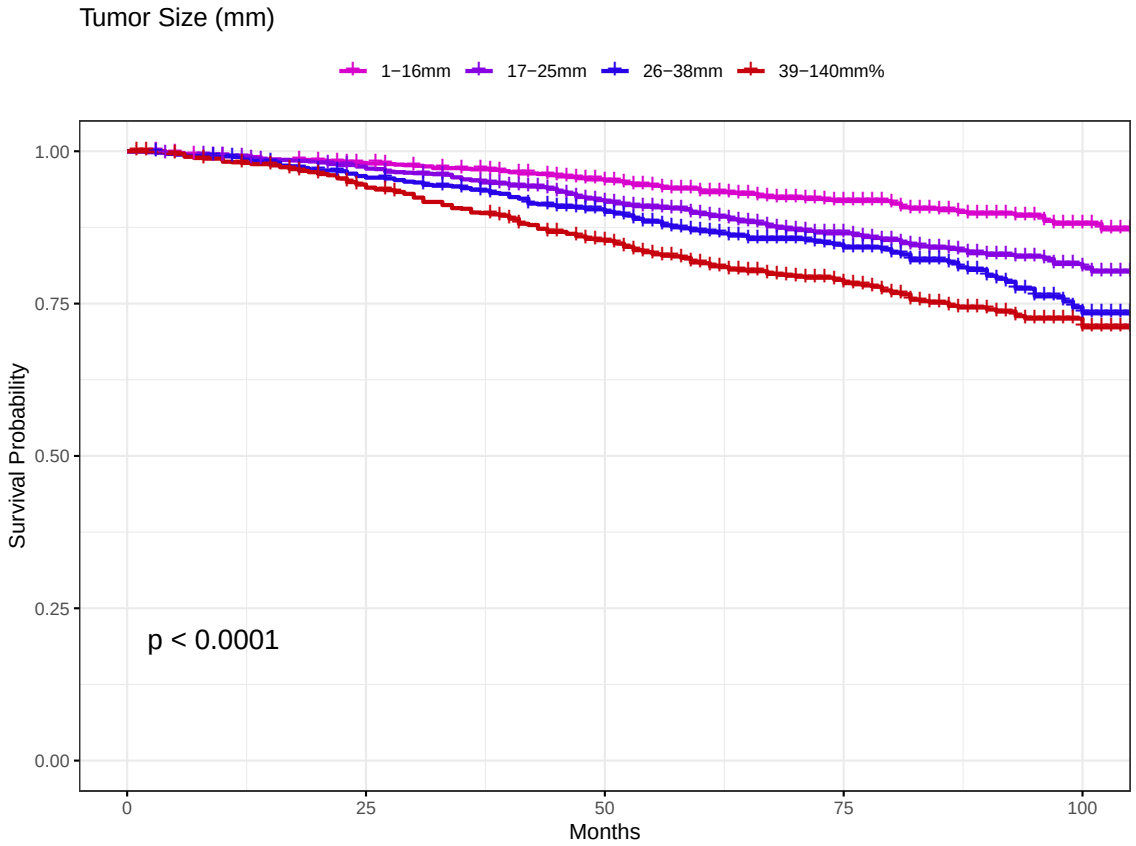


Figure 2. Kaplan-Meier survival curves for blood pressure binary variable (left), and age variable (right). The survival time for individuals with high blood pressure is significantly different from individuals with normal blood pressure. Survival time for individuals aged 70 appears significantly different than all other age groups. P-values for log rank tests do not specify which group(s) are significantly different, however plots can be used to discern visual differences.

```
# KMplots_positive
# KMplots_race
# KMplots_marital
# KMplots_t_stage
# KMplots_n_stage
# KMplots_6_stage
# KMplots_diff
# KMplots_grade
# KMplots_a_stage
# KMplots_estrogen
# KMplots_progesterone
```

Model Interpretation and Discussion

Conclusion

Reference

Appendix A:

Exploratory Analysis, Correlation Coefficient Plot, Kaplan-Meier plots with log rank test.

```
#load relevant packages
library(survminer)
library(survival)
library(ggplot2)
library(tidyverse)
library(My.stepwise)
library(corrr)
library(gtools)

#import data
cancer_df =
  read_csv("Project_2_data.csv") |>
  janitor::clean_names() |>
  as.data.frame()
```

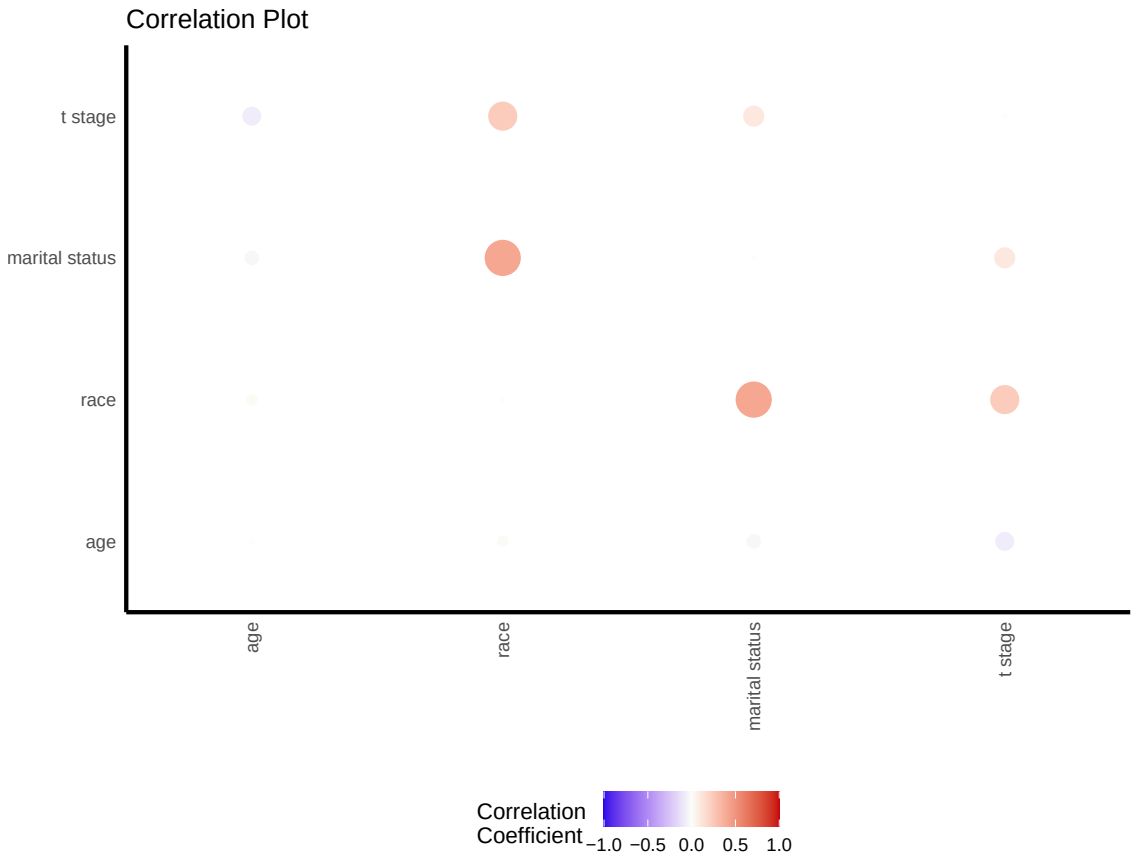
```
## Rows: 4024 Columns: 16
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (11): Race, Marital Status, T Stage, N Stage, 6th Stage, differentiate, ...
## dbl  (5): Age, Tumor Size, Regional Node Examined, Reginol Node Positive, Su...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Change the factor variables
col_names =
  cancer_df |> select_if(is.character) |> colnames() |> as.vector()

#variable summary. Output not shown
variablesummary <- lapply(cancer_df[,1:14], table)

# Change character variable "Status" to binary variables
cancer_df =
  cancer_df |>
  mutate(
    status =
      case_match(
        status,
        "Alive" ~ 0,
        "Dead" ~ 1)
  )
```

```
#correlation plot
res.cor <- correlate(cancer_df[,1:14])
```

```
## Non-numeric variables removed from input: `race`, `marital_status`, `t_stage`, `n_stage`, `x6th_stage
## Correlation computed with
## * Method: 'pearson'
## * Missing treated using: 'pairwise.complete.obs'
```

```r
cancer_cor <- res.cor %>% stretch()
cancer_cor[is.na(cancer_cor)]<-0

cancer_cor %>%
  ggplot(aes(x, y, col=r)) +
  geom_tile(col="white", fill="white") +
  geom_point(aes(size = abs(r)), shape=16) +
  labs(x = "", y = "", col = "Correlation \nCoefficient", title="Correlation Plot") +
  scale_color_gradient2(low="#2903e8",high="#c7000d",mid="#fbfef9", limits=c(-1,1))+
  scale_x_discrete(expand = c(0,0),labels=c("age", "race", "marital status", "t stage", "n stage", "6th
  scale_y_discrete(expand = c(0,0), labels = c("age", "race", "marital status", "t stage", "n stage", "
  scale_size(range=c(1,8), guide = NULL)+
  theme(axis.line = element_line(colour = "black",
                    linewidth = 1, linetype = "solid"),
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```

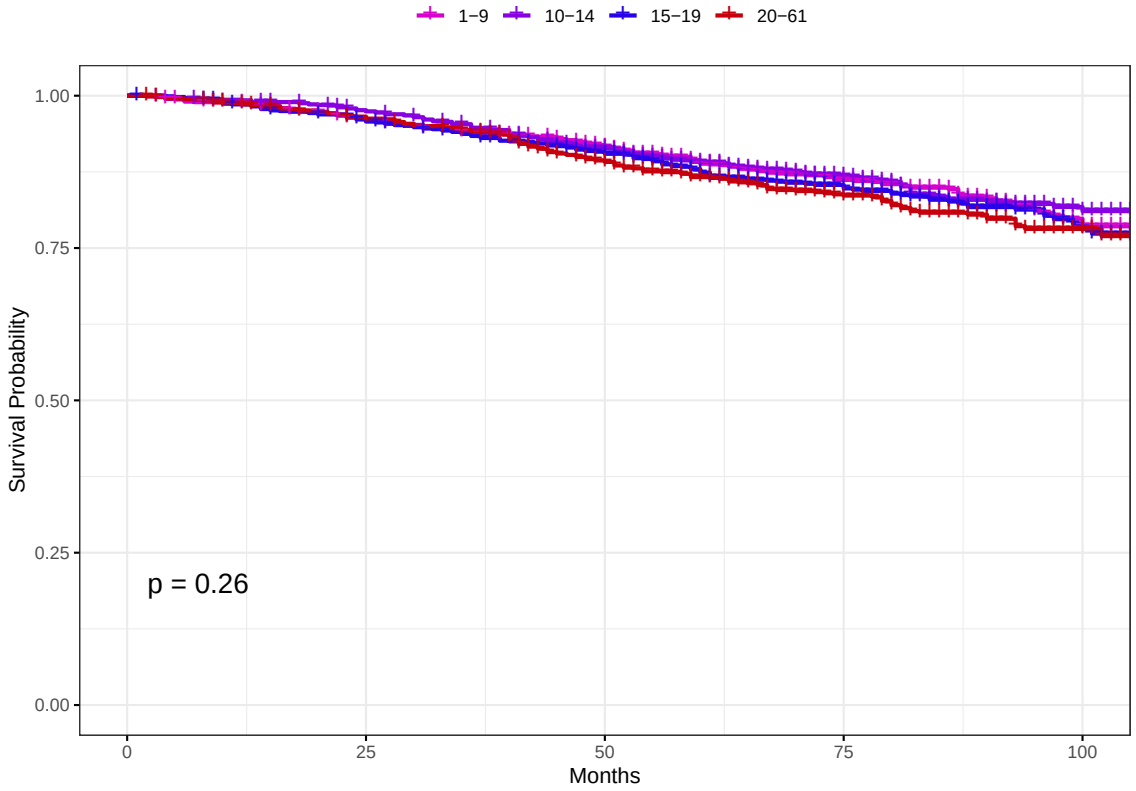



```r
#divide continuous variables into quartiles
cancer_df =
  cancer_df |>
  mutate(
    agequant = quantcut(age,4),
    tumor_size_quant = quantcut(tumor_size,4),
    regional_examined_quant = quantcut(regional_node_examined,4),
```

```
    reginol_positive_quant = quantcut(reginol_node_positive,4)
  )
```

```
#Kaplan-Meier estimators for significant variables. Plots found in Results section
KM_age <- survfit(Surv(survival_months, status)~agequant, data = cancer_df)
KM_tumor_size <- survfit(Surv(survival_months, status)~tumor_size_quant, data = cancer_df)
KM_reginol_positive <- survfit(Surv(survival_months, status)~reginol_positive_quant, data = cancer_df)
KM_race = survfit(Surv(survival_months, status)~race, data = cancer_df)
KM_marital = survfit(Surv(survival_months, status)~marital_status, data = cancer_df)
KM_t_stage = survfit(Surv(survival_months, status)~t_stage, data = cancer_df)
KM_n_stage = survfit(Surv(survival_months, status)~n_stage, data = cancer_df)
KM_6_stage = survfit(Surv(survival_months, status)~x6th_stage, data = cancer_df)
KM_diff = survfit(Surv(survival_months, status)~differentiate, data = cancer_df)
KM_grade = survfit(Surv(survival_months, status)~grade, data = cancer_df)
KM_a_stage = survfit(Surv(survival_months, status)~a_stage, data = cancer_df)
KM_estrogen = survfit(Surv(survival_months, status)~estrogen_status, data = cancer_df)
KM_progesterone = survfit(Surv(survival_months, status)~progesterone_status, data = cancer_df)

KMplots_age<- ggsurvplot(fit = KM_age, data = cancer_df, conf.int = F, title = "Age", ylab = "Survival

KMplots_tumor <- ggsurvplot(fit = KM_tumor_size, data = cancer_df, conf.int = F, title = "Tumor Size (m

KMplots_positive <- ggsurvplot(fit = KM_reginol_positive, data = cancer_df, conf.int = F, title = "Regi

KMplots_race<- ggsurvplot(fit = KM_race, data = cancer_df, conf.int = F, title = "Race", ylab = "Surviva

KMplots_marital<- ggsurvplot(fit = KM_marital, data = cancer_df, conf.int = F, title = "Marital Status"

KMplots_t_stage<- ggsurvplot(fit = KM_t_stage, data = cancer_df, conf.int = F, title = "T Stage: Adjust

KMplots_n_stage<- ggsurvplot(fit = KM_n_stage, data = cancer_df, conf.int = F, title = "N Stage: Adjust

KMplots_6_stage<- ggsurvplot(fit = KM_6_stage, data = cancer_df, conf.int = F, title = "6th Stage: Brea

KMplots_diff<- ggsurvplot(fit = KM_diff, data = cancer_df, conf.int = F, title = "Differentiate", ylab =

KMplots_grade<- ggsurvplot(fit = KM_grade, data = cancer_df, conf.int = F, title = "Grade", ylab = "Sur

KMplots_a_stage<- ggsurvplot(fit = KM_a_stage, data = cancer_df, conf.int = F, title = "A Stage", ylab =

KMplots_estrogen <- ggsurvplot(fit = KM_estrogen, data = cancer_df, conf.int = F, title = "Estrogen Sta

KMplots_progesterone <- ggsurvplot(fit = KM_progesterone, data = cancer_df, conf.int = F, title = "Prog
```

```
#Kaplain-Meier estimators and plots for non-significant variables
KM_regional_examined <- survfit(Surv(survival_months, status)~regional_examined_quant, data = cancer_df

KMplots_examined <- ggsurvplot(fit = KM_regional_examined, data = cancer_df, conf.int = F, title = "", y

KMplots_examined
```


1−9
10−14
15−19
20−61
1.00
0.75
0.50
0.25
0.00
Survival Probability
p = 0.26
0
25
50
75
100
Months