Breast cancer survival prediction

Yangyang Chen

2023-12-17

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*

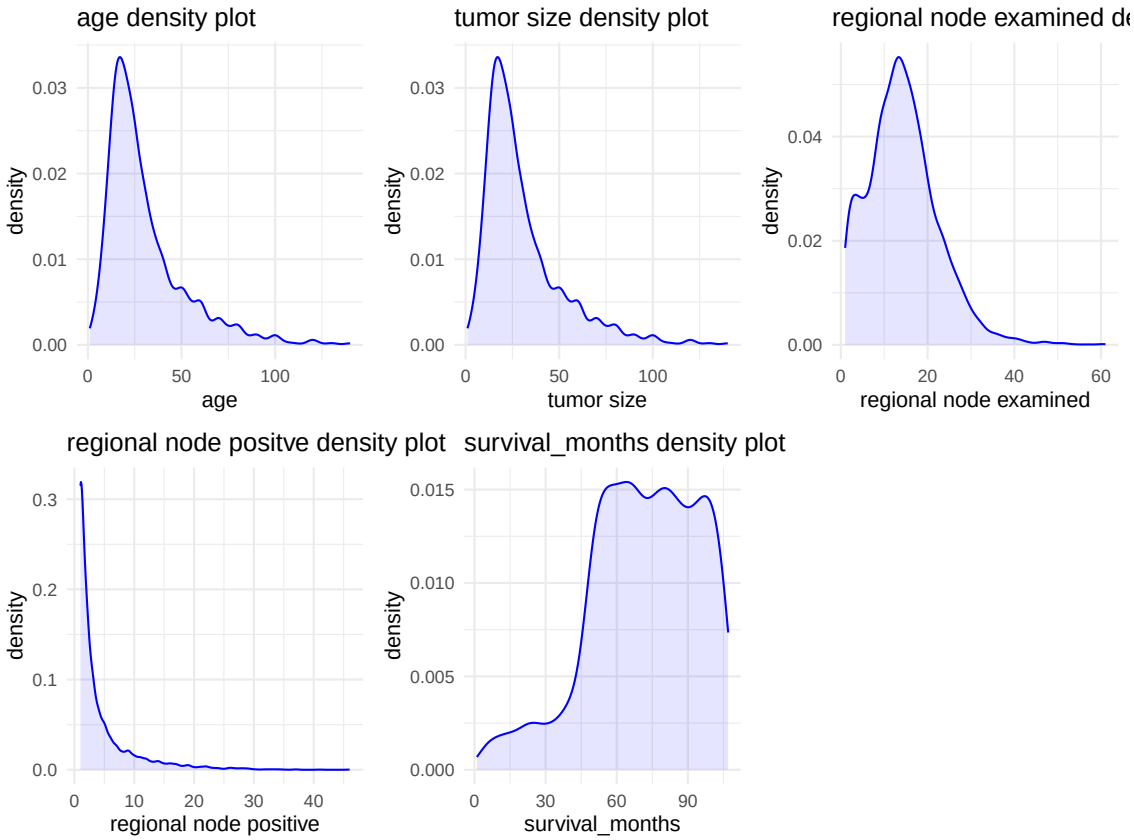
EDA

Methods

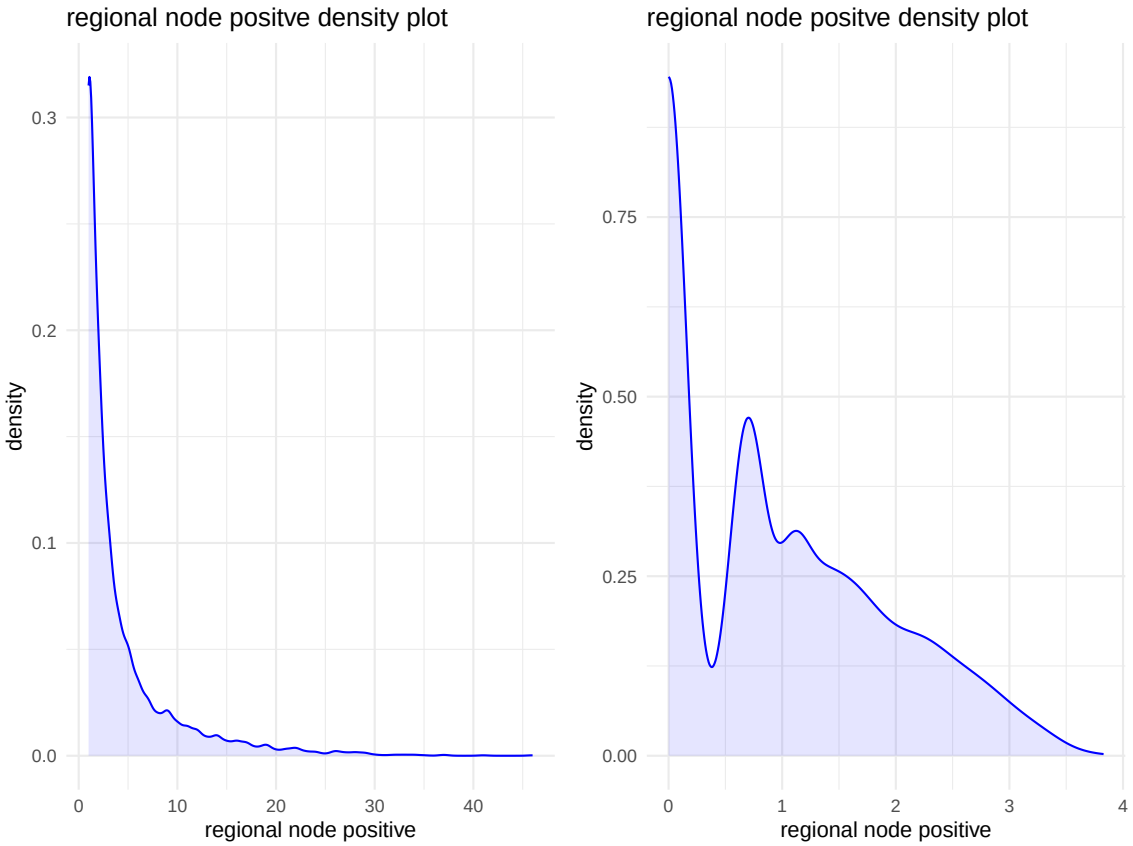
1. Data Description

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



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



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

Methods

The Kaplan-Meier estimator stands as one of the most extensively employed techniques for approximating survival functions in datasets featuring non-informative right censoring. It is utilized to perform univariate analysis on the impact of population groups within predictor variables on survival time. Also referred to as the product-limit estimator, the Kaplan-Meier method is non-parametric, eliminating the need to specify a distribution for survival time. Calculated as a point estimate for survival time at each recorded event time (such as each death in this context), the Kaplan-Meier estimator reflects the conditional probability of surviving at that particular time, and is written as

$$\hat{S}(t_i) = 1 - \frac{d_i}{n_i}$$

where

- $\hat{S}$ is the estimated survival probability,
- t_i is the event time,
- n_i is the number of individuals alive (at risk) right before t_i,
- d_i is the number of deaths at t_i. The cumulative Kaplan-Meier estimator for the survival function S(t) at all recorded event times in the data is the product of all $\hat{S}$ point estimates (Liu, 2012)

$$S(t) = \prod_{i=1}^{j} \left(1 - \frac{d_i}{n_i}\right)$$

where

- j is the total number of recorded events,
- d_i is the number of deaths at time t_i,
- n_i is the number of individuals at risk just before time t_i.

The computation of the Kaplan-Meier estimator is constrained to instances of recorded event times, excluding consideration of censoring times. However, the impact of censoring on outcomes remains salient, as individuals subject to censorship are removed from the cohort of individuals extant immediately prior to a designated event time t_i. Within Kaplan-Meier plots, alterations in the survival curve manifest exclusively at event times, with censoring times demarcated by tick marks along the survival curve.

To discern differentials in survival time contingent upon predictor variables, these variables are categorized into discrete population groups. Binary variables facilely lend themselves to dichotomous classification, characterized by values of 0 or 1. The Kaplan-Meier estimator is then applied to compute the survival function for each categorical group. Conversely, continuous variables are stratified into four populations, aligning with respective quartiles of the dataset. Kaplan-Meier plots serve as a visual apparatus to elucidate distinctions in survival functions among variable populations and to pinpoint potential non-proportionalities within survival curves.

Non-weighted log-rank tests are invoked to identify statistically significant disparities between Kaplan-Meier survival functions. This methodology, being non-parametric, circumvents reliance on the precise temporal occurrences of events, emphasizing instead the ordinal sequencing of events. The assignment of ranks is contingent upon the chronological sequence of events, with the initial recorded death accorded a rank of 1, followed sequentially by subsequent mortalities. It is imperative to underscore that log-rank tests are unsuitable for deployment within a multivariate milieu, their utility being confined to the comparison of disparities within specified populations—either binary or quartile groupings—pertaining to each variable.

In the context of the log-rank test for comparing two survival functions, the null hypothesis posits equivalence between the survival function of the first group and that of the second group. Conversely, the alternative hypothesis posits statistical non-equivalence in the survival functions. This comparative analysis is denoted as the two-sample log-rank test.

$$H_0 : S(t_1) = S(t_2)$$

$$H_A : S(t_1) \neq S(t_2)$$

The null hypothesis for log rank tests of more than two survival curves is that every survival curve in the test is equal. The alternative hypothesis is that at least one of the survival curves is different. The log-rank test does not specify which curve is the different one. Hypotheses for continuous variables, grouped into four quartiles, have four survival survival functions to test, and will be labeled as the four sample log-rank test (LaMorte, 2016)

$$H_0 : S(t_1) = S(t_2) = S(t_3) = S(t_4)$$

$$H_A : \text{At least one survival function } S(t_i) \text{ is not equal to the other survival functions}$$

To calculate the log rank test, the number of observed events is the sum of all events recorded at each time t (LaMorte, 2016). The number of expected events is defined as

$$E_{jt} = N_{jt} * \frac{O_t}{N_t}$$

where

- E_{jt} = Expected number of events at time t in sample j
- N_{jt} = Number of people at risk right before time t in sample j
- O_t = Number of Observed events across all samples
- N_t = Number of people at risk right before time t across all samples.

The log rank test statistic is calculated as a summation of the difference between the number observed events and expected events at each time interval, for each sample (LaMorte, 2016).

$$\chi^2 = \sum \left(\frac{(\sum O_{jt} - \sum E_{jt})^2}{\sum E_{jt}} \right)$$

where

- χ^2 is the log-rank test statistic on a chi-square distribution
- O_{jt} is the number of observed events for the jth group over time
- E_{jt} is the number of expected events for the jth group over time

The log-rank test statistic follows a chi-square distribution on $k-1$ degrees of freedom, where k is the number of groups. A two-sample log-rank test will have 1 degree of freedom ($k = 2$), and a four-sample log-rank test will have 3 degrees of freedom ($k = 4$). The significance, or p-value of the test can be found in a table of Critical Values or through calculation in R using the `pchisq()` function. For this analysis, a significance level of $\alpha = 0.05$ is used to determine test significance. P-values greater than 0.05 fail to reject the null hypothesis H_0, and p-values less than 0.05 reject the null hypothesis in favor of the alternative hypothesis. In this report, p-values calculated with log rank test are displayed on the Kaplan-Meier plots for each significant variable in the Results section, and for all other variables in Appendix A. Log-rank test significance will help determine which variables should be included in a multivariate model, and indicate which variables are significant to survival time in a univariate setting.

Log-rank tests were non-weighted because there is no evidence in the data description that survival differences are more meaningful if they occur closer to the beginning of the study, or closer to the end.

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.

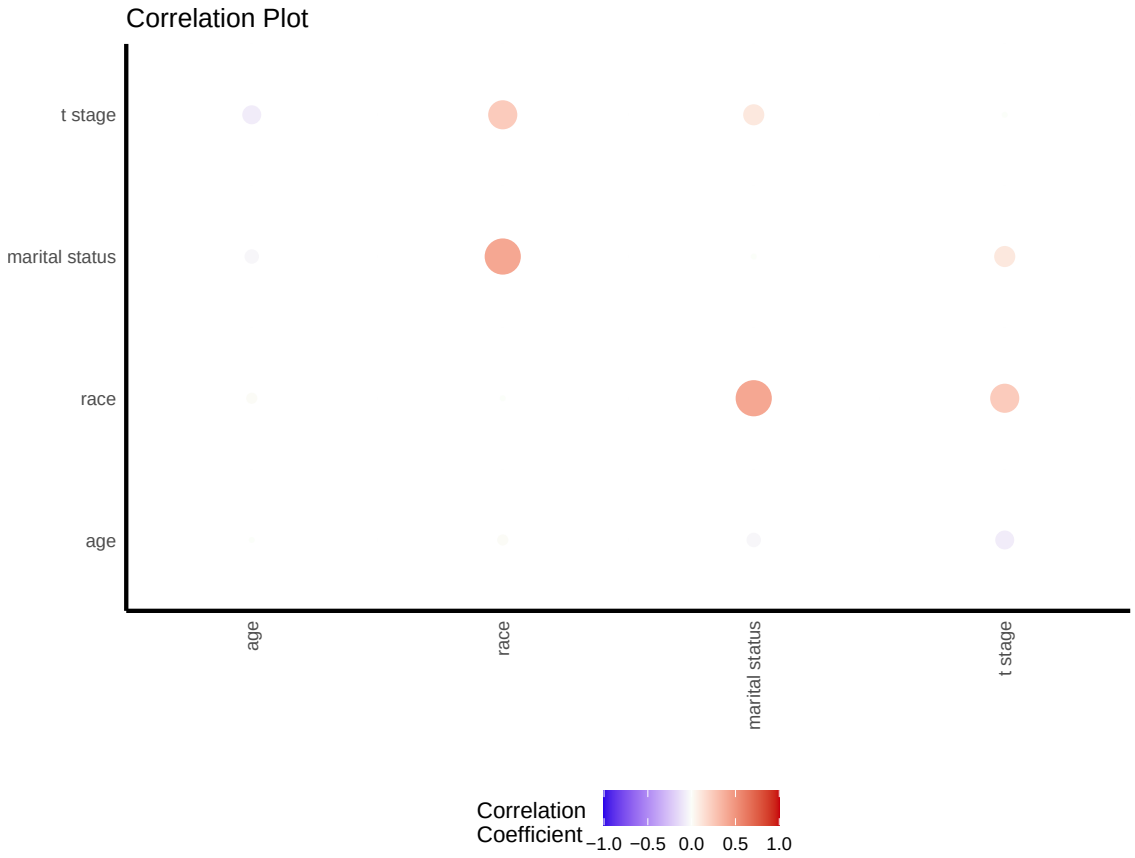
Model Interpretation and Discussion

Conclusion

Reference

Appendix A:

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


Correlation Plot
t stage
marital status
race
age
age
race
marital status
t stage
Correlation
Coefficient
−1.0 −0.5 0.0 0.5 1.0

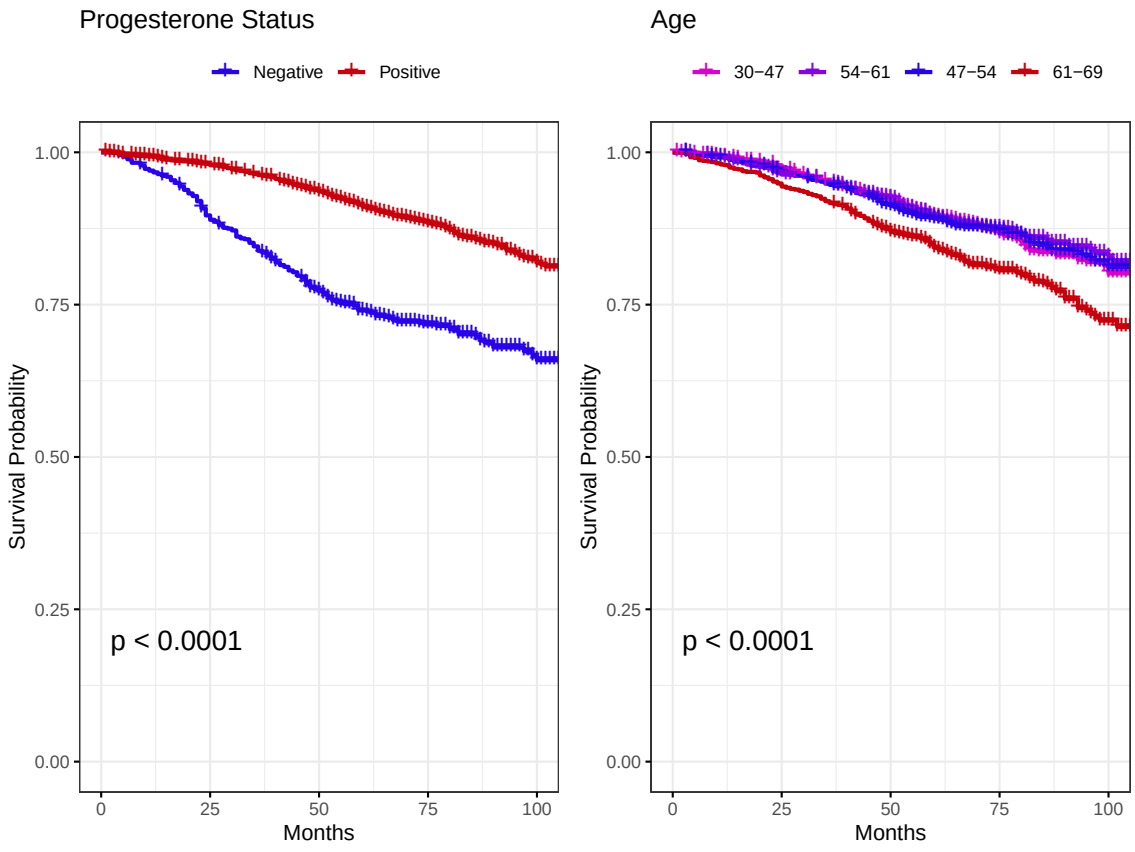

Progesterone Status
Negative
Positive
1.00
0.75
0.50
0.25
0.00
Survival Probability
p < 0.0001
0
25
50
75
100
Months
Age
30−47
54−61
47−54
61−69
1.00
0.75
0.50
0.25
0.00
Survival Probability
p < 0.0001
0
25
50
75
100
Months

Figure 2. Kaplan-Meier survival curves for progesterone binary variable (left), and age variable (right). The survival time for individuals with positive progesterone is significantly different from individuals with negative progesterone. Survival time for individuals aged 61-69 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.

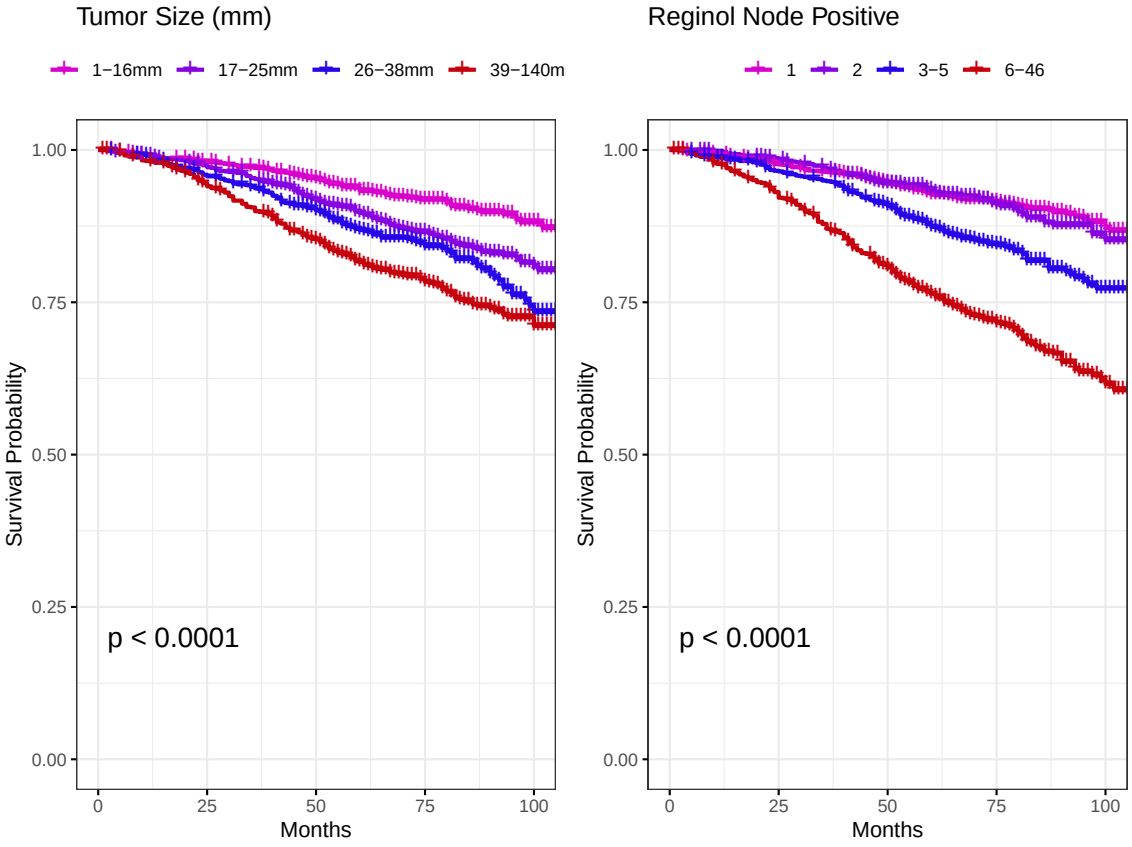


Figure 3. Kaplan-Meier survival curves for tumor size (left) and reginol node positive (right). The survival time for individuals with tumor size larger than 39mm appear significantly different from all other groups. The group with the smallest tumor size 1-16mm does not have the highest survival probability of the four groups over time. Reginol node positive, measured in number, appears to have multiple significant differences between groups. The group with the lowest positive node has the highest survival percentage over time, and the group with the highest positive node has the lowest survival percentage over time.

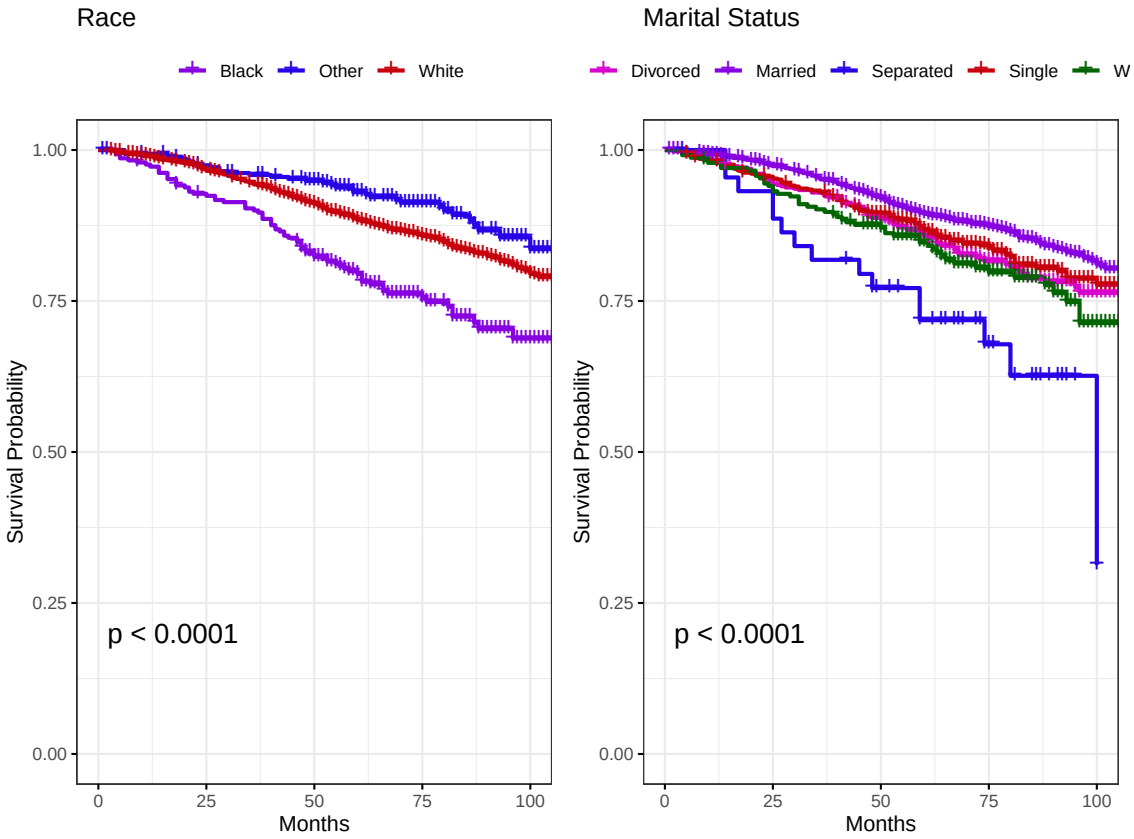


Figure 3. Kaplan-Meier survival curves for race (left) and Marital Status (right). The survival time for individuals with race larger than 39mm appear significantly different from all other groups. The group with the smallest tumor size 1-16mm does not have the highest survival probability of the four groups over time. Reginol node positive, measured in number, appears to have multiple significant differences between groups. The group with the lowest positive node has the highest survival percentage over time, and the group with the highest positive node has the lowest survival percentage over time.

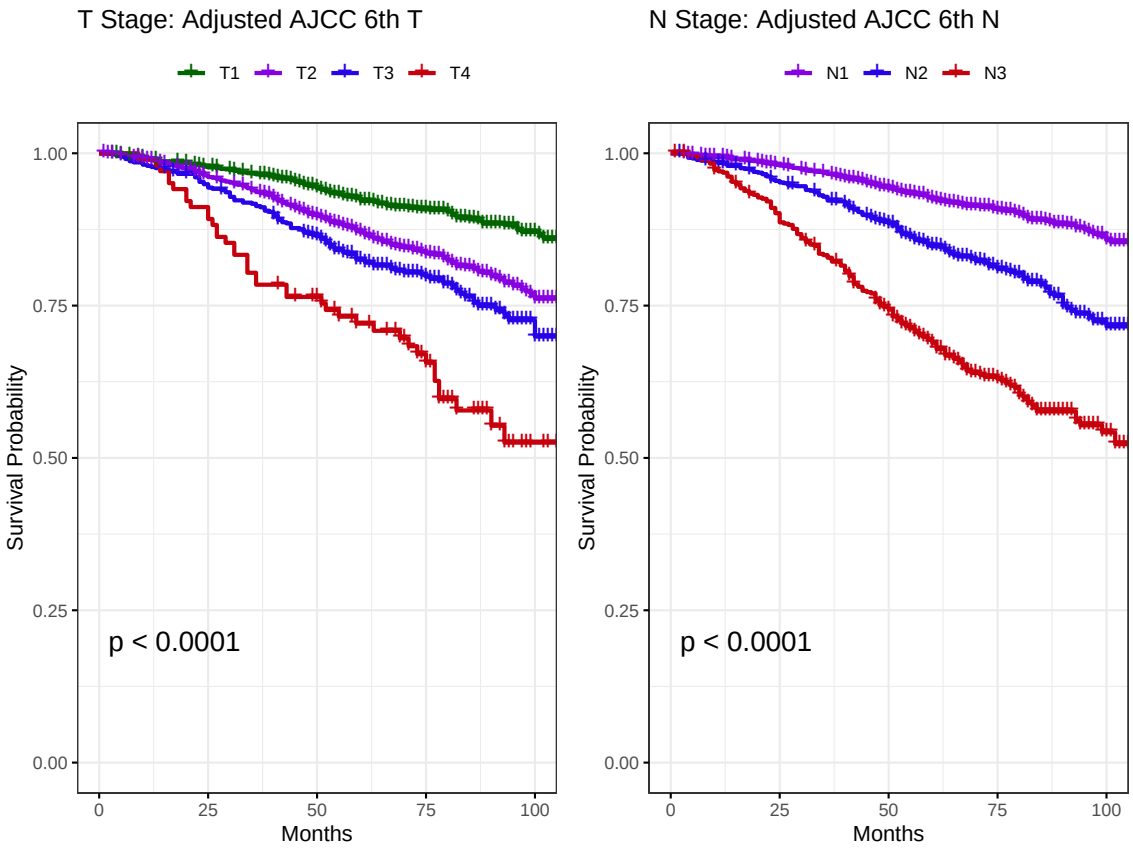


Figure 4. Kaplan-Meier survival curves for N Stage (left) and serum creatinine level (right). The survival time for individuals with ejection fraction percentages less than 30% appear significantly different from all other groups. The group with the highest ejection fraction percentage (45-80%) does not have the highest survival probability of the four groups over time. Serum creatinine, measured in mg/dL, appears to have multiple significant differences between groups. The group with the lowest levels has the highest survival percentage over time, and the group with the highest levels has the lowest survival percentage over time.

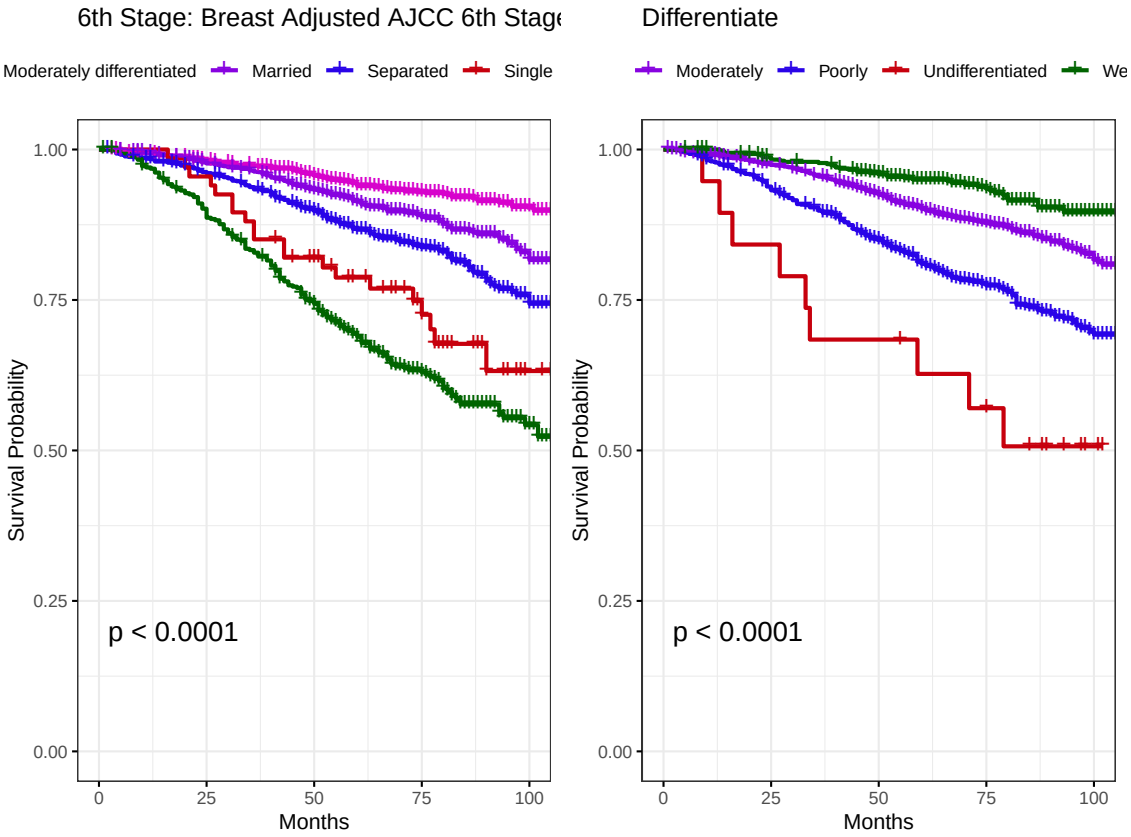


Figure 3. Kaplan-Meier survival curves for Breast Adjusted AJCC 6th Stage, multiple levels variable (left), and Differentiate variable (right). The survival time for widows is significantly different from married individuals. Survival time for undifferentiated individuals appears significantly different than all other differentiate 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.

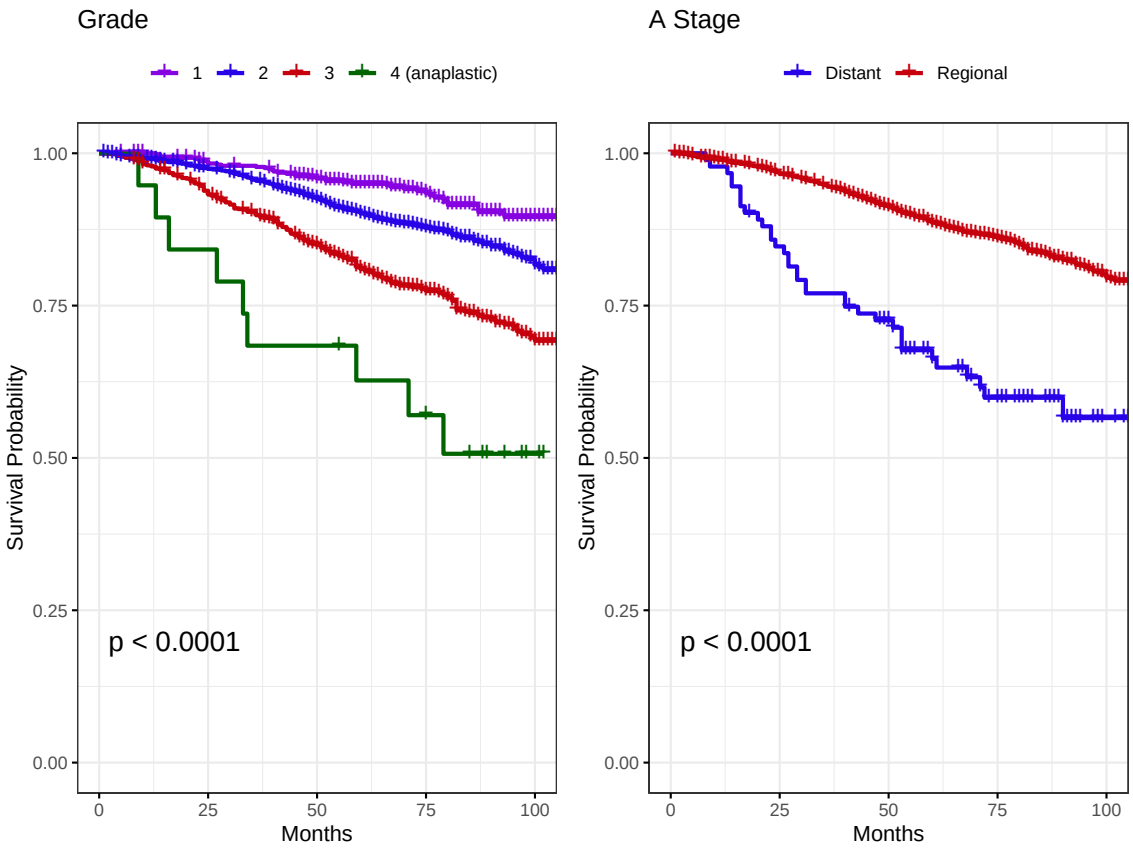


Figure 4. Kaplan-Meier survival curves for grade (left) and a stage (right). The survival time for individuals with grade 4 (anaplastic) appear significantly different from all other groups. The group with the highest grade 1 does have the highest survival probability of the four groups over time. A Stage, measured in binary level, appears to have two significant differences between groups. The group with the regional stage has the highest survival percentage over time, and the group with the distant stage has the lowest survival percentage over time.

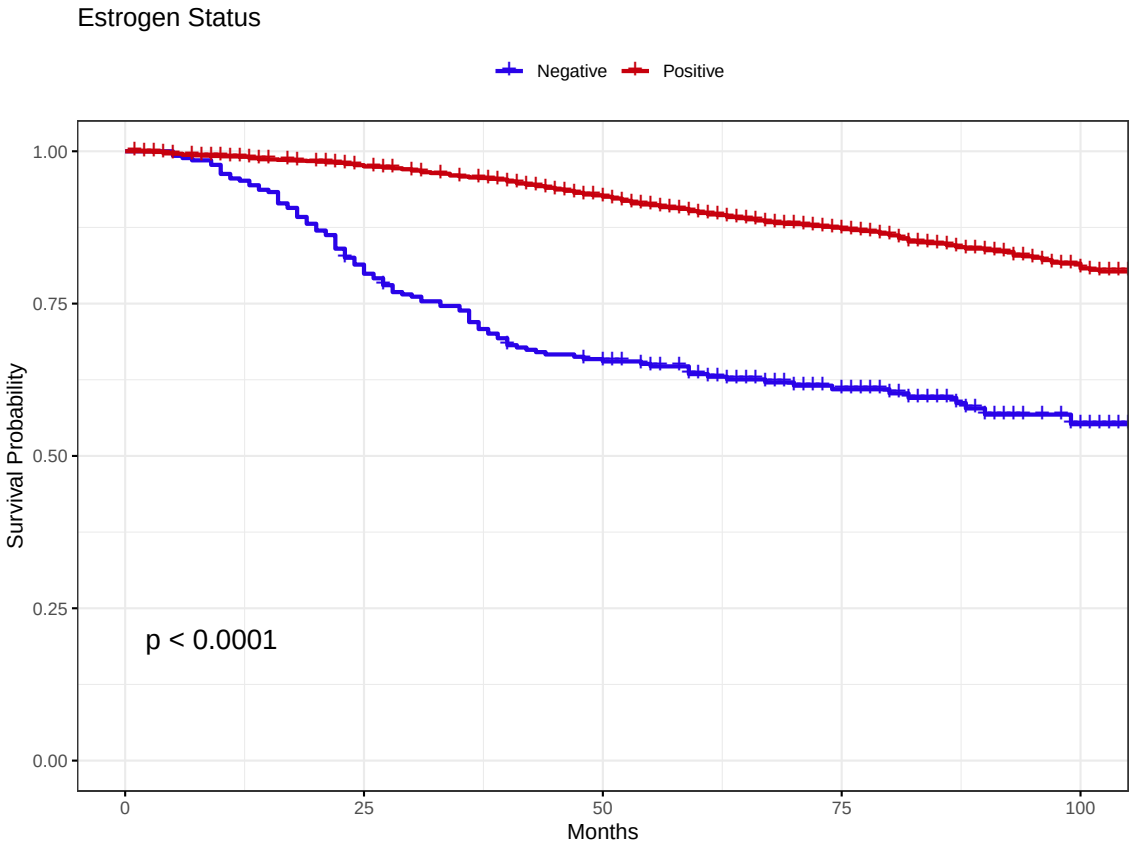


Figure 5. Kaplan-Meier survival curves of each Estrogen Status group. The survival time for individuals with negative Estrogen Status appears significantly different, and less than positive groups.