

# Data analysis

12/09/2023

## Libraries

```
library(tidyverse)
library(readr)
library(boot)
library(table1)
library(gridExtra)
library(MASS)
library(car)
library(dplyr)
library(leaps)
library(corrplot)
```

## Data Clean

```
breastcancer_data =
  read_csv("Project_2_data.csv") |>
  janitor::clean_names()

## # 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.

summary(breastcancer_data)

##      age          race        marital_status       t_stage
##  Min.   :30.00   Length:4024   Length:4024   Length:4024
##  1st Qu.:47.00   Class :character  Class :character  Class :character
##  Median :54.00   Mode  :character  Mode  :character  Mode  :character
##  Mean   :53.97
##  3rd Qu.:61.00
##  Max.   :69.00
##      n_stage        x6th_stage       differentiate       grade
##  Length:4024   Length:4024   Length:4024   Length:4024
```

```

##  Class :character  Class :character  Class :character  Class :character
##  Mode  :character  Mode  :character  Mode  :character  Mode  :character
##
##
##
##    a_stage          tumor_size      estrogen_status  progesterone_status
##  Length:4024        Min.   : 1.00  Length:4024        Length:4024
##  Class :character  1st Qu.: 16.00  Class :character  Class :character
##  Mode  :character  Median : 25.00  Mode  :character  Mode  :character
##                      Mean   : 30.47
##                      3rd Qu.: 38.00
##                      Max.   :140.00
##  regional_node_examined  reginol_node_positive survival_months
##  Min.   : 1.00        Min.   : 1.000        Min.   : 1.0
##  1st Qu.: 9.00        1st Qu.: 1.000        1st Qu.: 56.0
##  Median :14.00        Median : 2.000        Median : 73.0
##  Mean   :14.36        Mean   : 4.158        Mean   : 71.3
##  3rd Qu.:19.00        3rd Qu.: 5.000        3rd Qu.: 90.0
##  Max.   :61.00        Max.   :46.000        Max.   :107.0
##    status
##  Length:4024
##  Class :character
##  Mode  :character
##
##
##
##bc = breastcancer_data |>
#mutate(
#  race=case_when(
#    race == "White" ~ 1,
#    race == "Black" ~ 2,
#    race == "Other" ~ 3),
#  marital_status=case_when(
#    marital_status == "Married" ~ 1,
#    marital_status == "Divorced" ~ 2,
#    marital_status == "Single" ~ 3,
#    marital_status == "Widowed" ~ 4,
#    marital_status == "Separated" ~ 5),
#  t_stage=case_when(
#    t_stage == "T1" ~ 1,
#    t_stage == "T2" ~ 2,
#    t_stage == "T3" ~ 3,
#    t_stage == "T4" ~ 4),
#  n_stage=case_when(
#    n_stage == "N1" ~ 1,
#    n_stage == "N2" ~ 2,
#    n_stage == "N3" ~ 3),
#  x6th_stage=case_when(
#    x6th_stage == "IIA" ~ 1,
#    x6th_stage == "IIIA" ~ 2,
#    x6th_stage == "IIIC" ~ 3,
#    x6th_stage == "IIB" ~ 4,
#    x6th_stage == "IIIB" ~ 5),

```

```

differentiate=case_when(
  differentiate == "Poorly differentiated" ~ 1,
  differentiate == "Moderately differentiated" ~ 2,
  differentiate == "Well differentiated" ~ 3,
  differentiate == "Undifferentiated" ~ 4),
grade=case_when(
  grade == "1" ~ 1,
  grade == "2" ~ 2,
  grade == "3" ~ 3,
  grade == "anaplastic; Grade IV" ~ 4),
a_stage=case_when(
  a_stage == "Regional" ~ 1,
  a_stage == "Distant" ~ 0),
estrogen_status=case_when(
  estrogen_status == "Positive" ~ 1,
  estrogen_status == "Negative" ~ 0),
progesterone_status=case_when(
  progesterone_status == "Positive" ~ 1,
  progesterone_status == "Negative" ~ 0),
status=case_when(
  status == "Alive" ~ 1,
  status == "Dead" ~ 0)
)

```

## Descriptive statistics for all variables

```

summary(bc)

##      age          race   marital_status     t_stage
##  Min. :30.00    Min. :1.000  Min. :1.000  Min. :1.000
##  1st Qu.:47.00  1st Qu.:1.000  1st Qu.:1.000  1st Qu.:1.000
##  Median :54.00  Median :1.000  Median :1.000  Median :2.000
##  Mean   :53.97  Mean   :1.231  Mean   :1.646  Mean   :1.785
##  3rd Qu.:61.00  3rd Qu.:1.000  3rd Qu.:2.000  3rd Qu.:2.000
##  Max.   :69.00  Max.   :3.000  Max.   :5.000  Max.   :4.000
##      n_stage      x6th_stage   differentiate      grade
##  Min. :1.000  Min. :1.000  Min. :1.000  Min. :1.000
##  1st Qu.:1.000 1st Qu.:1.000  1st Qu.:1.000  1st Qu.:2.000
##  Median :1.000  Median :2.000  Median :2.000  Median :2.000
##  Mean   :1.438  Mean   :2.405  Mean   :1.868  Mean   :2.151
##  3rd Qu.:2.000 3rd Qu.:4.000  3rd Qu.:2.000  3rd Qu.:3.000
##  Max.   :3.000  Max.   :5.000  Max.   :4.000  Max.   :4.000
##      a_stage      tumor_size   estrogen_status  progesterone_status
##  Min.   :0.0000  Min.   : 1.00  Min.   :0.0000  Min.   :0.0000
##  1st Qu.:1.0000  1st Qu.: 16.00  1st Qu.:1.0000  1st Qu.:1.0000
##  Median :1.0000  Median : 25.00  Median :1.0000  Median :1.0000
##  Mean   :0.9771  Mean   : 30.47  Mean   :0.9332  Mean   :0.8265
##  3rd Qu.:1.0000  3rd Qu.: 38.00  3rd Qu.:1.0000  3rd Qu.:1.0000
##  Max.   :1.0000  Max.   :140.00  Max.   :1.0000  Max.   :1.0000
##      regional_node_examined reginol_node_positive survival_months     status
##  Min.   : 1.00        Min.   : 1.000        Min.   : 1.0   Min.   :0.0000

```

```

## 1st Qu.: 9.00      1st Qu.: 1.000      1st Qu.: 56.0      1st Qu.:1.0000
## Median :14.00      Median : 2.000      Median : 73.0      Median :1.0000
## Mean   :14.36      Mean   : 4.158      Mean   : 71.3      Mean   :0.8469
## 3rd Qu.:19.00      3rd Qu.: 5.000      3rd Qu.: 90.0      3rd Qu.:1.0000
## Max.   :61.00      Max.   :46.000      Max.   :107.0      Max.   :1.0000

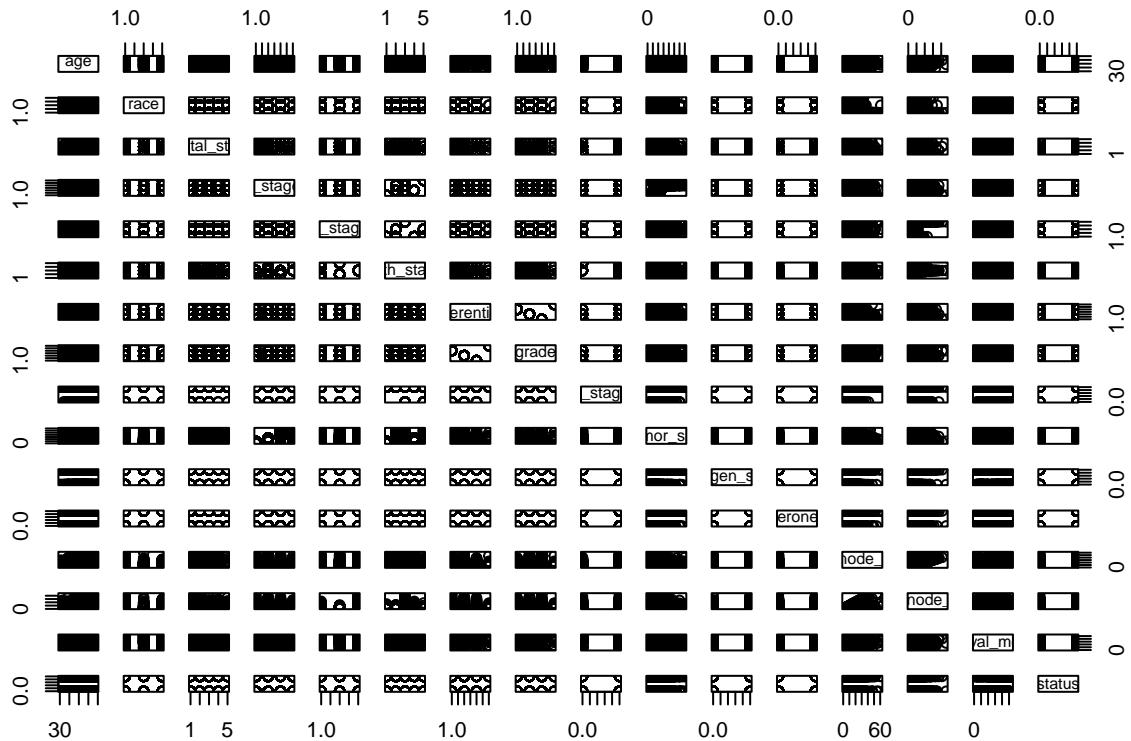
```

We change race, marital\_status, t\_stage, n\_stage, x6th\_stage, differentiate, and grade into multiple numeric levels, while a\_stage, estrogen\_status, progesterone\_status, and status to binary levels. The above variables are categorical variables.

And age, tumor\_size, regional\_node\_examined, reginol\_node\_positive, and survival\_months are numeric variables.

## Covariance and Correlation

```
plot(bc)
```



```
cor(bc) |>
knitr::kable(digits=4,caption="Correlation for all variables")
```

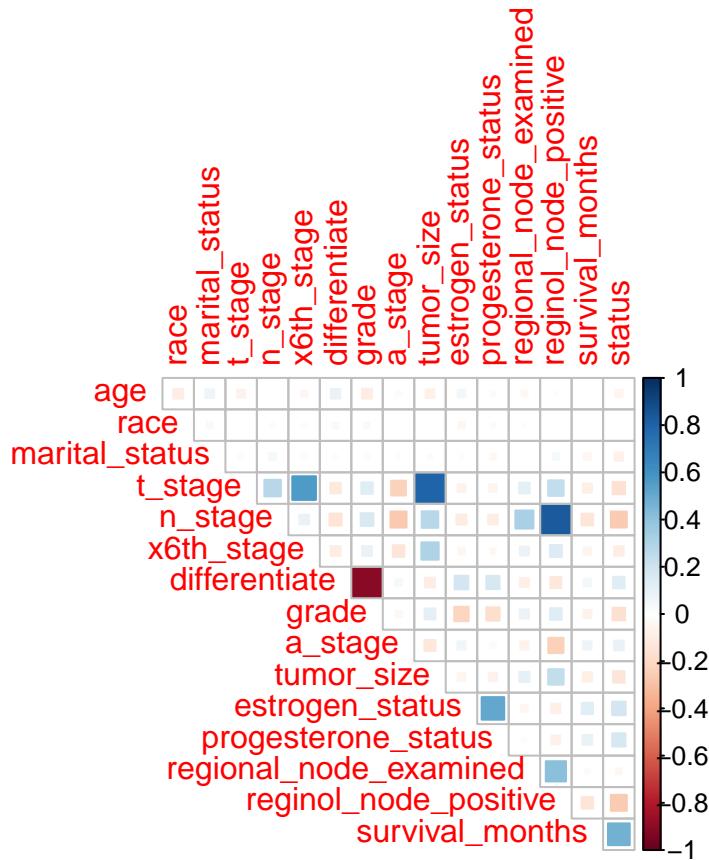
Table 1: Correlation for all variables

	age	race	marital_st	t_stag	n_stag	x6th_stag	differentiate	grade	estrogen_status	progesterone_status	regional_node_examined	reginol_node_positive	tumor_size	survival_months	status
age	1.0000	-0.0755	-	0.0029	-	0.0932	-	0.0209	-	0.0598	-	-	-	0.0126	-

	age	race	marital_status	t_stage	6th_stage	differentiate	a_stage	estrogen_size	progesterone_size	regional_node	status	node_size	oleic_acid	survival_months		
race	-	1.0000	0.0349	0.0036	0.0190	0.0173	-	0.0301	0.0071	-	-	0.0113	0.0090	-	0.0042	
		0.0970					0.0336		0.0020			0.0425	0.0209		0.0025	
marital_status	0.0755	0.0349	0.0000	0.0250	0.0450	0.0221	-	0.0206	0.0207	-	-	0.0077	0.0443	-	-	
							0.0117		0.0174			0.0198	0.0357		0.0487 0.0731	
t_stage	-	0.0036	0.0257	1.0000	0.2770	0.5637	-	0.1315	0.8092	-	-	0.1141	0.2431	-	-	
		0.0669					0.1102		0.2211			0.0610	0.0576		0.0857 0.1547	
n_stage	0.0020	0.0190	0.0457	0.2770	0.0000	0.0939	-	0.1625	0.2779	-	-	0.3283	0.8381	-	-	
							0.1488		0.2606			0.1020	0.0937		0.1396 0.2558	
x6th_stage	0.0173	0.0221	0.5637	0.0930	0.0000	-	0.0972	0.3034	-	-	0.0826	0.1427	-	-		
		0.0450					0.0999		0.1372			0.0417	0.0309		0.0536 0.0919	
differentiate	0.0932	-	-	-	-	1.0000	-	0.0437	-	0.1868	0.1758	-	-	0.0584	0.1342	
							0.0336	0.0117	0.1102	0.1488	0.0999	0.9083	0.0995		0.0834 0.1229	
grade	-	0.0300	0.0206	0.1316	0.1626	0.0972	-	1.0000	-	0.1194	-	-	0.0844	0.1353	-	-
		0.0993					0.9083		0.0395			0.2113	0.1799		0.0677 0.1614	
a_stage	0.0209	-	-	-	-	0.0437	-	1.0000	-	0.0656	0.0265	-	-	0.0701	0.0966	
							0.0020	0.0174	0.2210	0.2606	0.1372	0.0395	0.1239		0.0690 0.2328	
tumor_size	0.0070	0.0207	0.8090	0.2770	0.3034	-	0.1194	-	1.0000	-	-	0.1044	0.2423	-	-	
		0.0772					0.0995		0.1239			0.0596	0.0699		0.0869 0.1342	
estrogen_size	0.0508	-	-	-	-	0.1868	-	0.0656	-	1.0000	0.5133	-	-	0.1285	0.1847	
							0.0425	0.0198	0.0610	0.1020	0.0417	0.2113	0.0596		0.0448 0.0860	
progesterone_status	-	-	-	-	0.1758	-	0.0265	-	0.5133	1.0000	-	-	-	0.0960	0.1771	
							0.0210	0.0200	0.0357	0.0570	0.0930	0.0309	0.1799	0.0699		0.0181 0.0781
regional_node	0.0118	0.0077	0.1140	0.3280	0.0826	-	0.0844	-	0.1044	-	-	1.0000	0.4116	-	-	
		0.0333					0.0834		0.0690			0.0448	0.0181		0.0221 0.0348	
reginol_size	0.0044	0.0090	0.0443	0.2430	0.8380	0.1427	-	0.1353	0.2423	-	-	0.4116	1.0000	-	-	
							0.1229		0.2328			0.0860	0.0781		0.1352 0.2566	
survival_months	-	-	-	-	0.0584	-	0.0701	-	0.1285	0.0960	-	-	-	1.0000	0.4765	
							0.0094	0.0026	0.0487	0.0850	0.1390	0.0536	0.0677	0.0869		0.0221 0.1352
status	-	0.0042	-	-	-	0.1342	-	0.0966	-	0.1847	0.1771	-	-	-	0.4765 1.0000	
		0.0559					0.0731	0.1540	0.2558	0.0919		0.1614	0.1342		0.0348 0.2566	

### Another plot for correlation

```
corrplot(cor(bc), method = "square", type = "upper", diag = FALSE)
```



## Exploratory visualisation

```
plot1age =
breastcancer_data|>
ggplot(aes(x = age)) +
geom_histogram(color = "blue", fill = alpha("blue", 0.5), binwidth = 5) +
theme_minimal() +
labs(
  title = "Age Distribution",
  x = "Age",
  y = "Frequency"
)
#plot1age
```

```
plot2race =
breastcancer_data|>
ggplot(aes(x = race)) +
geom_bar(color = "blue", fill = alpha("blue", 0.5))+
theme_minimal() +
labs(
  title = "Race Distribution",
  x = "Race",
  y = "Frequency"
)
```

```
#plot2race
```

```
plot3marital =
breastcancer_data|>
ggplot(aes(x = marital_status)) +
geom_bar(color = "blue", fill = alpha("blue", 0.5))+
theme_minimal() +
labs(
  title = "Marital Status Distribution",
  x = "Marital Status",
  y = "Frequency"
)
```

```
#plot3marital
```

```
plot4tstage =
breastcancer_data|>
ggplot(aes(x = t_stage)) +
geom_bar(color = "blue", fill = alpha("blue", 0.5))+
theme_minimal() +
labs(
  title = "T Stage Distribution",
  x = "T Stage",
  y = "Frequency"
)
```

```
#plot4tstage
```

```
plot5nstage =
breastcancer_data|>
ggplot(aes(x = n_stage)) +
geom_bar(color = "blue", fill = alpha("blue", 0.5))+
theme_minimal() +
labs(
  title = "N Stage Distribution",
  x = "N Stage",
  y = "Frequency"
)
```

```
#plot5nstage
```

```
plot6x6thstage =
breastcancer_data|>
ggplot(aes(x = x6th_stage)) +
geom_bar(color = "blue", fill = alpha("blue", 0.5))+
theme_minimal() +
labs(
  title = "x6th Stage Distribution",
  x = "x6th Stage",
  y = "Frequency"
)
```

```
#plot6x6thstage
```

```

plot7differentate =
breastcancer_data|>
ggplot(aes(x = differentiate)) +
geom_bar(color = "blue", fill = alpha("blue", 0.5))+ 
theme_minimal() +
labs(
  title = "Differentiate Distribution",
  x = "Differentiate Group",
  y = "Frequency"
)

```

*#plot7differentate*

```

plot8grade =
breastcancer_data|>
ggplot(aes(x = grade)) +
geom_bar(color = "blue", fill = alpha("blue", 0.5))+ 
theme_minimal() +
labs(
  title = "Grade Distribution",
  x = "Grade",
  y = "Frequency"
)

```

*#plot8grade*

```

plot9astage =
breastcancer_data|>
ggplot(aes(x = a_stage)) +
geom_bar(color = "blue", fill = alpha("blue", 0.5))+ 
theme_minimal() +
labs(
  title = "A_stage Distribution",
  x = "A Stage",
  y = "Frequency"
)

```

*#plot9astage*

```

plot10tumorsize =
breastcancer_data|>
ggplot(aes(x = tumor_size)) +
geom_histogram(color = "blue", fill = alpha("blue", 0.5))+ 
theme_minimal() +
labs(
  title = "Tumor Size Distribution",
  x = "Tumor Size",
  y = "Frequency"
)

```

*#plot10tumorsize*

```

plot11estrogen =
breastcancer_data|>
ggplot(aes(x = estrogen_status)) +
geom_bar(color = "blue", fill = alpha("blue", 0.5))+ 
theme_minimal() +
labs(
  title = "Estrogen Status Distribution",
  x = "Estrogen Status",
  y = "Frequency"
)

```

#plot11estrogen

```

plot12progesterone =
breastcancer_data|>
ggplot(aes(x = progesterone_status)) +
geom_bar(color = "blue", fill = alpha("blue", 0.5))+ 
theme_minimal() +
labs(
  title = "Progesterone Status Distribution",
  x = "Progesterone Status",
  y = "Frequency"
)

```

#plot12progesterone

```

plot13nodeexamined =
breastcancer_data|>
ggplot(aes(x = regional_node_examined)) +
geom_histogram(color = "blue", fill = alpha("blue", 0.5))+ 
theme_minimal() +
labs(
  title = "Regional Node Examined Distribution",
  x = "Examined Regional Node",
  y = "Frequency"
)

```

#plot13nodeexamined

```

plot14nodepositive =
breastcancer_data|>
ggplot(aes(x = reginol_node_positive)) +
geom_histogram(color = "blue", fill = alpha("blue", 0.5))+ 
theme_minimal() +
labs(
  title = "Regional Node Positive Distribution",
  x = "Positive Reginol Node",
  y = "Frequency"
)

```

#plot14nodepositive

Y1 = survive months; (numeric)

```

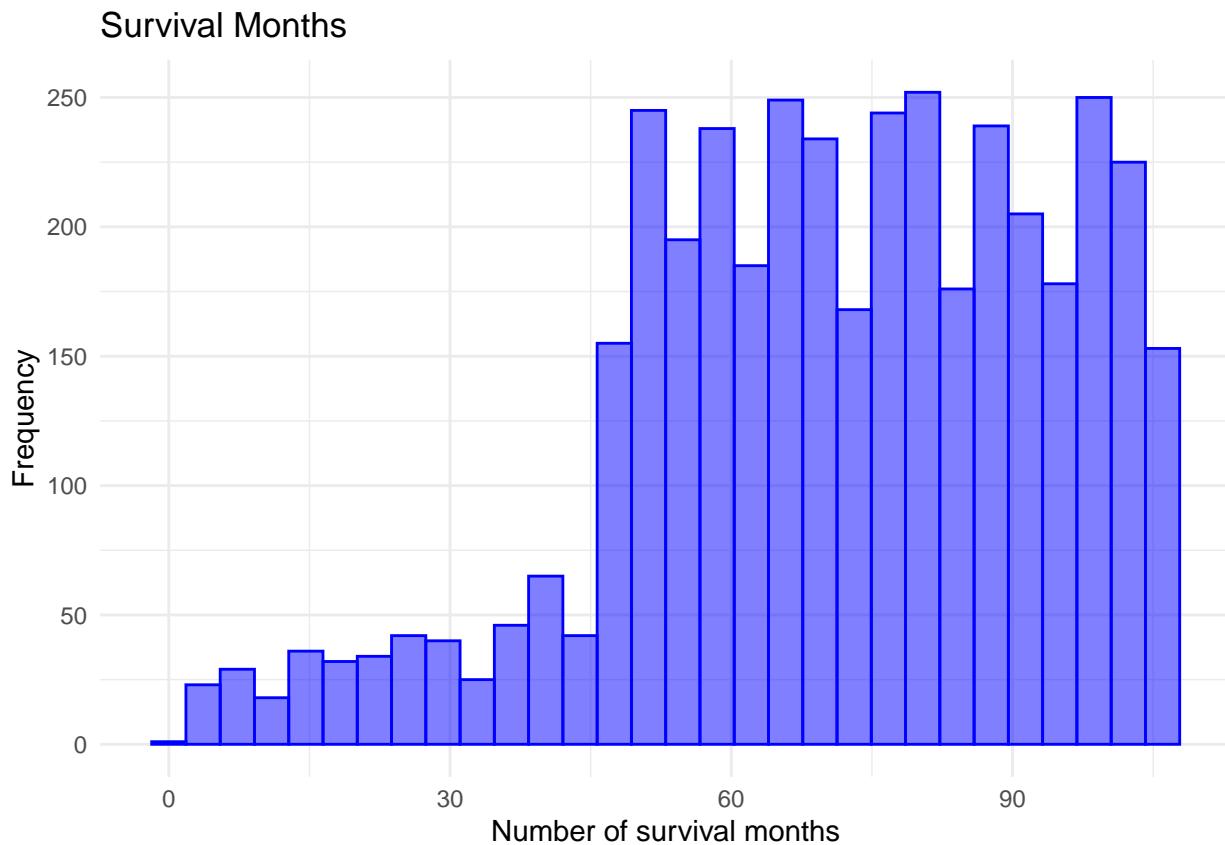
Y2 = status; (binary)

plot15survivalmonths =
breastcancer_data|>
ggplot(aes(x = survival_months)) +
geom_histogram(color = "blue", fill = alpha("blue", 0.5))+ 
theme_minimal() +
labs(
  title = "Survival Months",
  x = "Number of survival months",
  y = "Frequency"
)

plot15survivalmonths

```

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

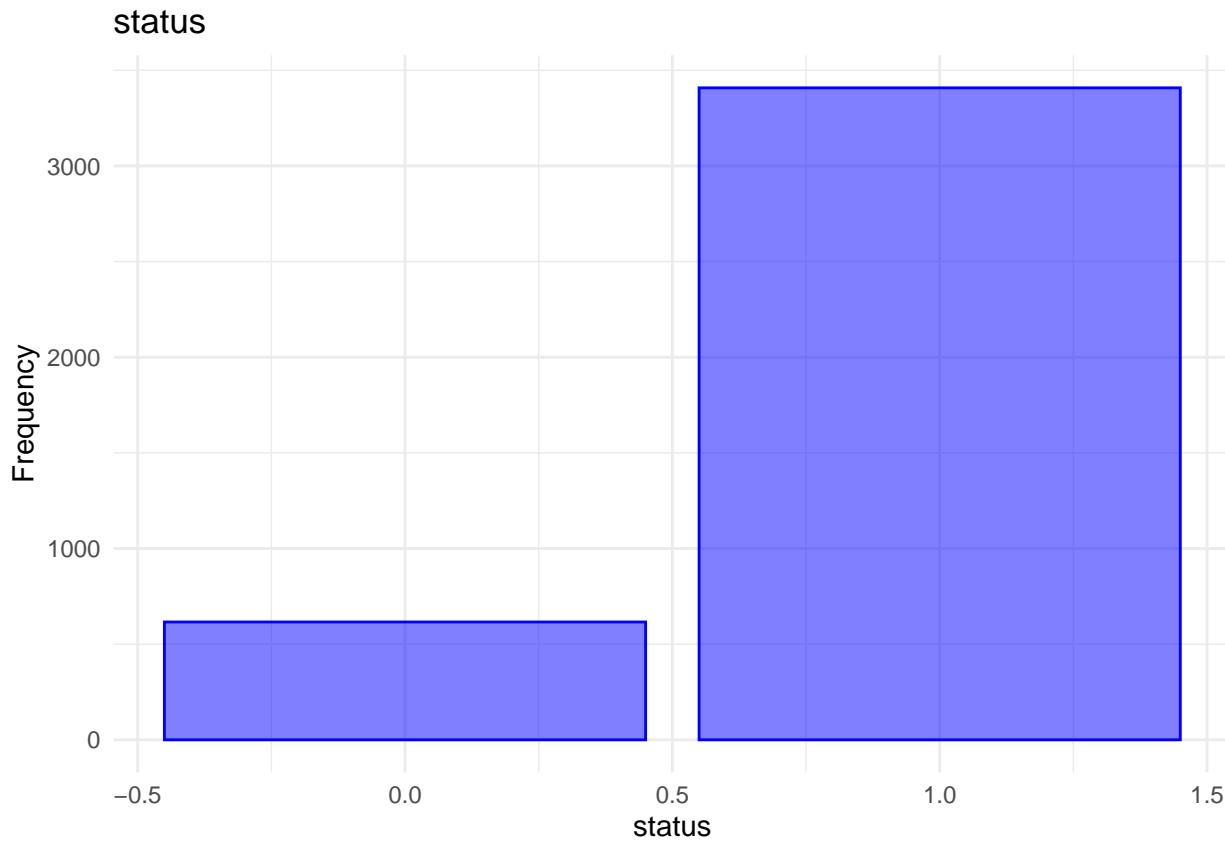


```

plot16status =
bc|>
ggplot(aes(x = status)) +
geom_bar(color = "blue", fill = alpha("blue", 0.5))+ 
theme_minimal() +
labs(
  title = "status",
  x = "status",
  y = "Frequency"
)

```

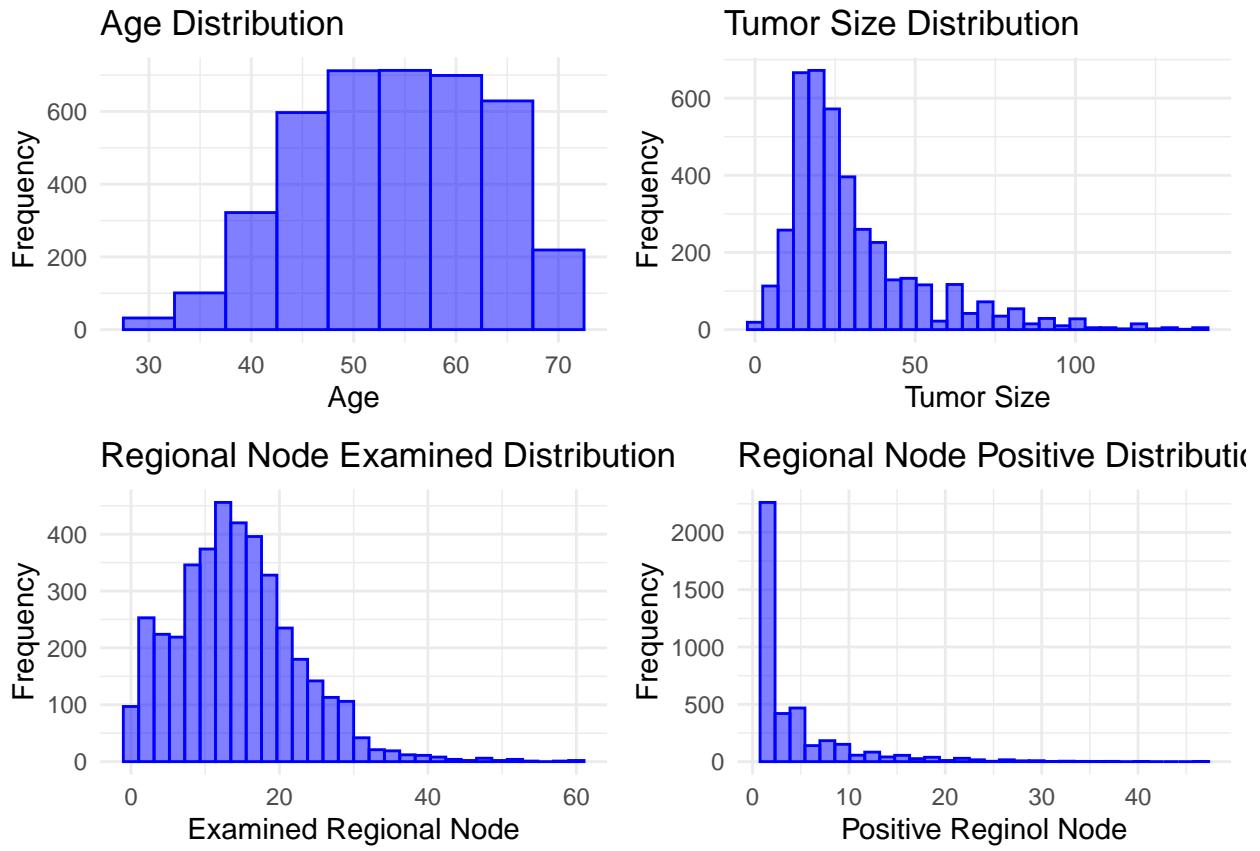
```
)  
plot16status
```



### Summarized plots for covariates

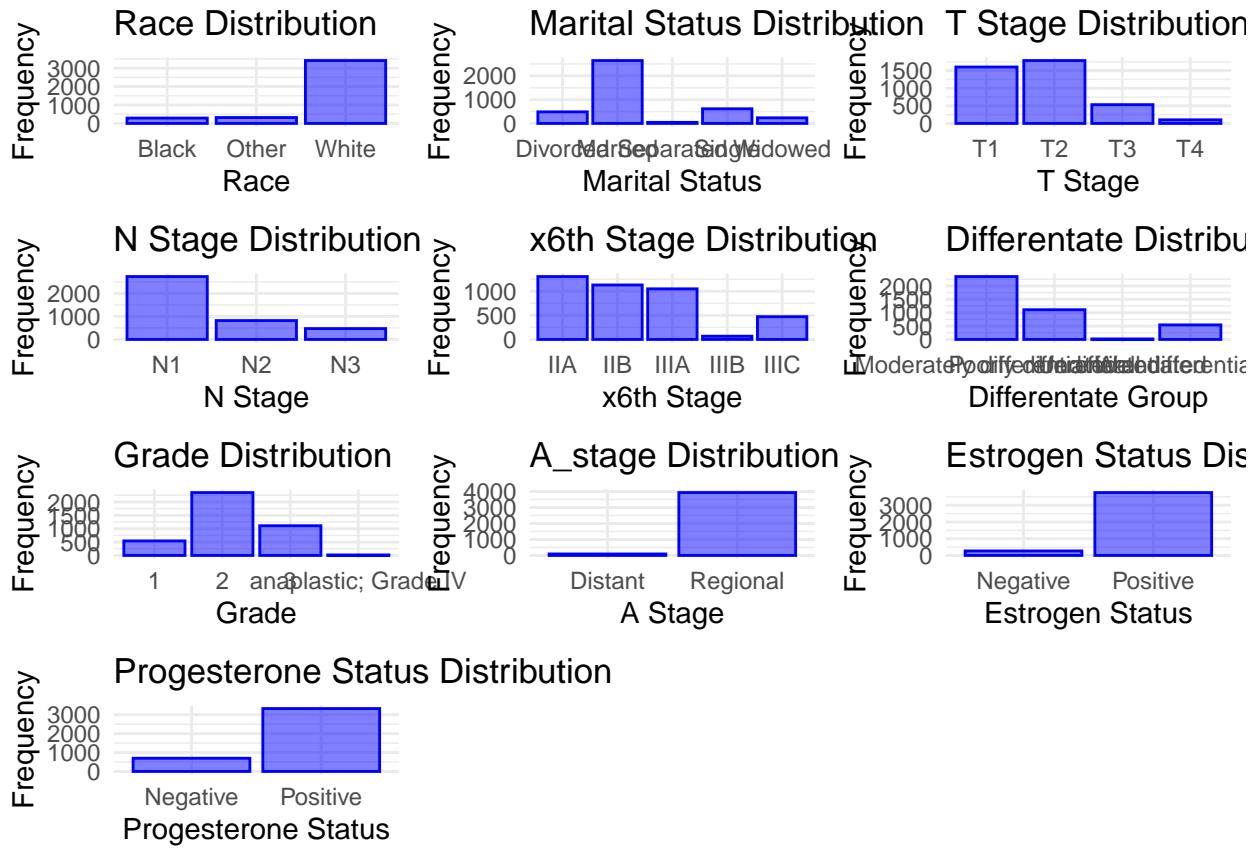
```
grid.arrange(plot1age, plot10tumorsize, plot13nodeexamined,  
            plot14nodepositive, ncol = 2)
```

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.  
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.  
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```



We can see that age is approximately normal, while tumor size, regional node examined, and regional node positive are skewed.

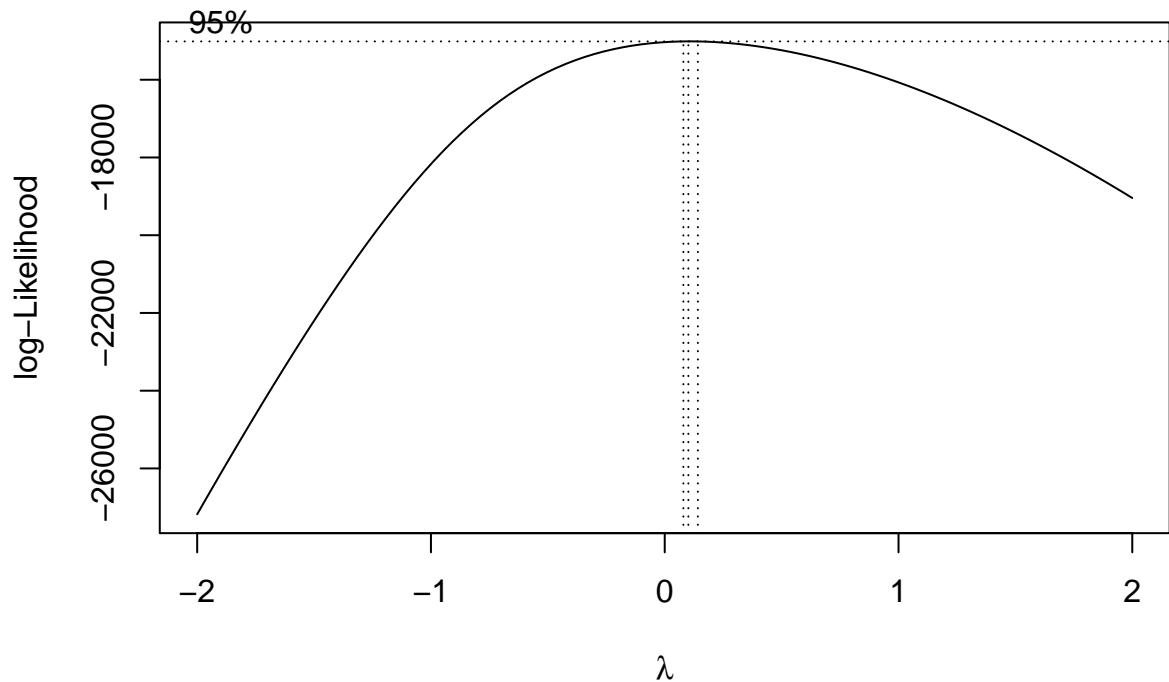
```
grid.arrange(plot2race, plot3marital,plot4tstage,
            plot5nstage, plot6x6thstage,plot7differentate,
            plot8grade,plot9astage,plot11estrogen,plot12progesterone, ncol = 3)
```



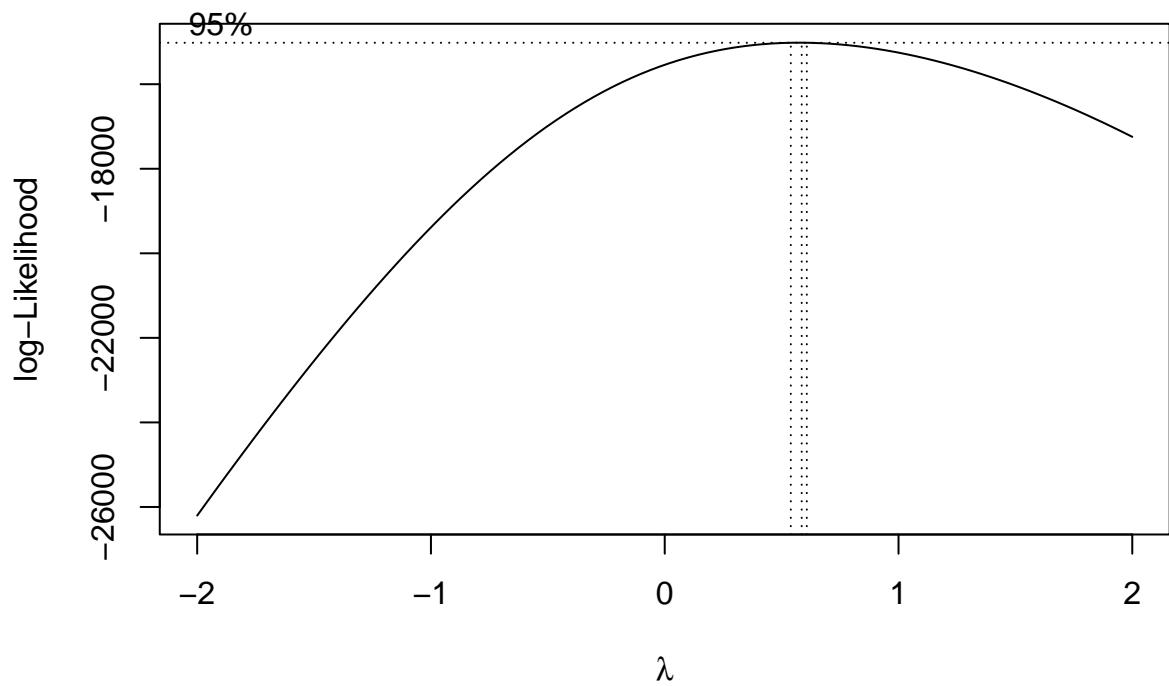
## Transformation

We know that the tumor size, regional node examined, and regional node positive are skewed. We should do transformation on these variables. Before the transformation, we can use the Box-Cox plot to check which transformation work the best for them.

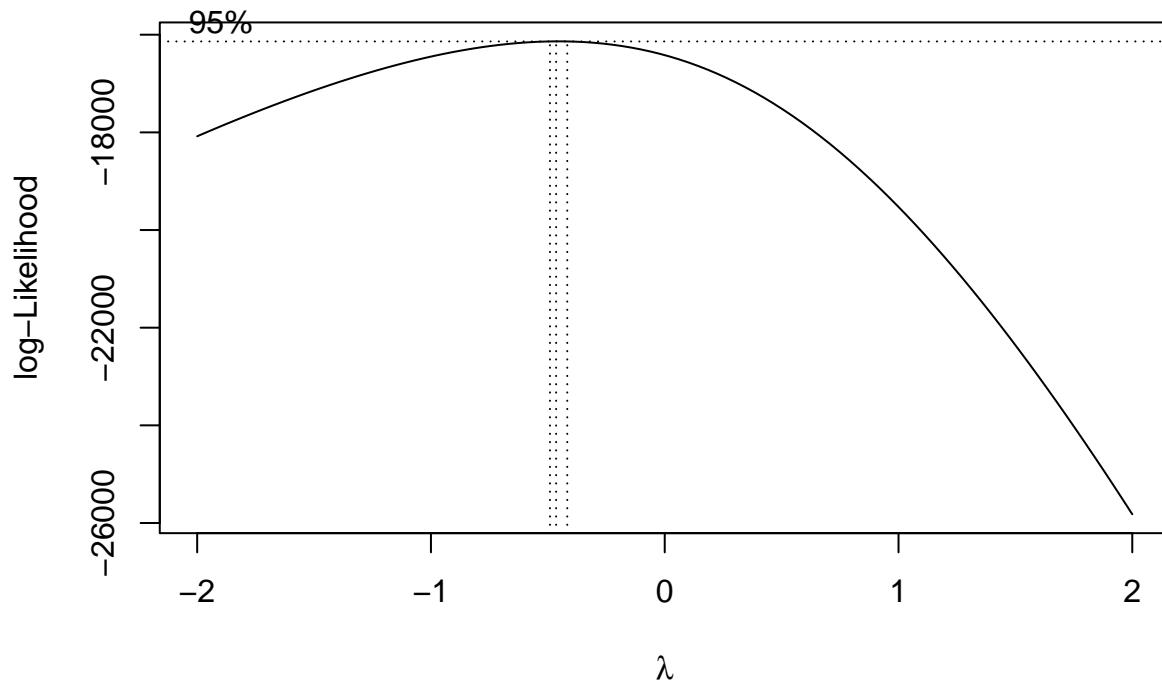
```
bc_transform_tumorsize <- boxcox(breastcancer_data$tumor_size ~ 1, lambda = seq(-2, 2, by=0.1))
```



```
bc_transform_regionalnode_examined <- boxcox(breastcancer_data$regional_node_examined ~ 1, lambda = seq(-2, 2, 0.01))
```



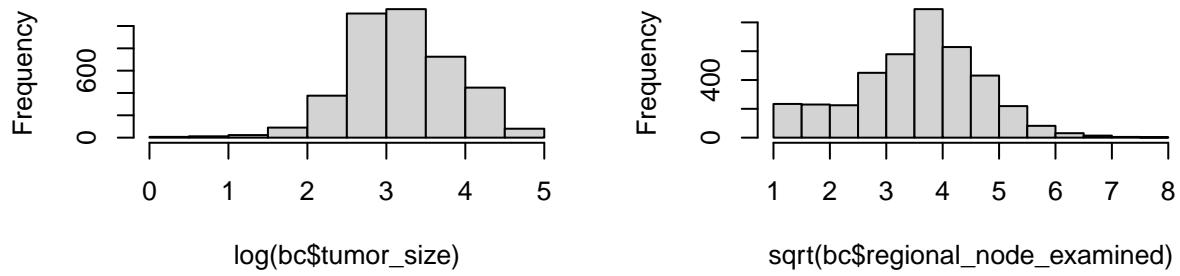
```
bc_transform_regionalnode_pos <- boxcox(breastcancer_data$regional_node_positive ~ 1, lambda = seq(-2, 2, 0.01))
```



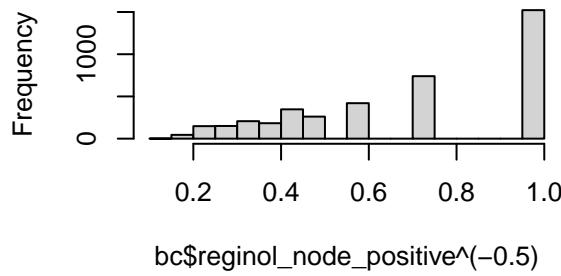
The lambda value of tumor size is close to 0, so we should use log transformation, while the lambda value of regional node examined is around 0.5, we should take a square root to the value, and the lambda value of regional node positive is around -0.5, so we should take a take square root and take an (-1) exponent for transformation.

```
par(mfrow = c(2, 2))
hist(log(bc$tumor_size))
hist(sqrt(bc$regional_node_examined))
hist(bc$reginol_node_positive**(-0.5))
```

## Histogram of log(bc\$tumor\_size) | histogram of sqrt(bc\$regional\_node\_examined)



## histogram of bc\$reginol\_node\_positive^(



We can see that tumor size and regional node examined become approximately normal after log transformation, while the regional node positive is still extremely skewed. Therefore, we may consider not using the variable of reginol\_node\_positive.

## Transformation model

```
newbc = bc |>
  mutate(ln_tumor=log(tumor_size),
        sqrt_examined=sqrt(regional_node_examined)) |>
  dplyr::select(-tumor_size) |>
  dplyr::select(-regional_node_examined) |>
  dplyr::select(-survival_months)
newbc

## # A tibble: 4,024 x 15
##       age   race marital_status t_stage n_stage x6th_stage differentiate grade
##     <dbl> <dbl>      <dbl>    <dbl>    <dbl>      <dbl>          <dbl> <dbl>
## 1     68     1          1        1        1          1            1     3
## 2     50     1          1        1        2          2            2     2
## 3     58     1          2        3        3          3            3     2
## 4     58     1          1        1        1          1            1     3
## 5     47     1          1        2        1          4            1     3
## 6     51     1          3        1        1          1            2     2
## 7     51     1          1        1        1          1            3     1
## 8     40     1          1        2        1          4            2     2
## 9     40     1          2        4        3          3            1     3
## 10    69     1          1        4        3          3            3     1
```

```

## # i 4,014 more rows
## # i 7 more variables: a_stage <dbl>, estrogen_status <dbl>,
## #   progesterone_status <dbl>, reginol_node_positive <dbl>, status <dbl>,
## #   ln_tumor <dbl>, sqrt_examined <dbl>

```

## chi-square tests

```

# Create an empty data frame to store chi-square test results
chi_square_results <- data.frame(Variable = character(),
                                 ChiSquare = numeric(),
                                 PValue = numeric(),
                                 stringsAsFactors = FALSE)

# Specify the dependent variable
dependent_var <- newbc$status

# List of independent variables
independent_vars <- newbc[,-13]

# Perform chi-square test for each independent variable
for (var in names(independent_vars)) {
  # Create a contingency table
  table <- table(dependent_var, newbc[[var]])

  # Perform chi-square test and suppress warnings for expected count less than 5
  test <- try(suppressWarnings(chisq.test(table))), silent = TRUE)

  # Store the results if the test was successful
  if (!inherits(test, "try-error")) {
    chi_square_results <- rbind(chi_square_results, data.frame(Variable = var,
                                                               ChiSquare = test$statistic,
                                                               PValue = test$p.value))
  } else {
    # Store NA if the test failed due to too many zero counts
    chi_square_results <- rbind(chi_square_results, data.frame(Variable = var,
                                                               ChiSquare = NA,
                                                               PValue = NA))
  }
}

chi_square_results

##                               Variable ChiSquare      PValue
## X-squared                  age  86.47154 1.880216e-05
## X-squared1                 race  27.97001 8.440929e-07
## X-squared2                marital_status  28.26381 1.102769e-05
## X-squared3                 t_stage 103.47631 2.779095e-22
## X-squared4                 n_stage 269.92914 2.430141e-59
## X-squared5                x6th_stage 281.64844 9.830332e-60
## X-squared6               differentiate 112.55628 3.091352e-24
## X-squared7                 grade 112.55628 3.091352e-24
## X-squared8                 a_stage  35.76473 2.226426e-09

```

```

## X-squared9      estrogen_status 135.15574 3.052608e-31
## X-squared10    progesterone_status 124.88539 5.392080e-29
## X-squared11    reginol_node_positive 343.65852 2.292691e-51
## X-squared12    ln_tumor 192.12355 1.560395e-06
## X-squared13    sqrt_examined 71.62539 4.501059e-02

```

Based on the above chi-squared table, each variable listed has been tested for independence with respect to the dependent variable, and each shows a significant relationship.

## Indicator Test

When y is status

```

# indicator test when y is status
categorical_vars <- c("race", "marital_status", "t_stage", "n_stage", "x6th_stage",
                      "differentiate", "grade", "a_stage",
                      "estrogen_status", "progesterone_status")

newbc[categorical_vars] <- lapply(newbc[categorical_vars], factor)

formula <- as.formula("status ~ race + marital_status + t_stage + n_stage + x6th_stage +
                        differentiate + grade + a_stage + estrogen_status + progesterone_status+ln_tumor")

model <- glm(formula, data = newbc, family = binomial())

summary(model)

```

```

##
## Call:
## glm(formula = formula, family = binomial(), data = newbc)
##
## Coefficients: (4 not defined because of singularities)
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)           1.826084   0.591231  3.089  0.00201 **
## race2                -0.517829   0.161866 -3.199  0.00138 **
## race3                 0.412988   0.202323  2.041  0.04123 *
## marital_status2     -0.208737   0.141798 -1.472  0.14100
## marital_status3     -0.137065   0.134860 -1.016  0.30946
## marital_status4     -0.226852   0.192395 -1.179  0.23836
## marital_status5     -0.870139   0.369406 -2.356  0.01850 *
## t_stage2              -0.241275   0.214882 -1.123  0.26151
## t_stage3              -0.463184   0.308144 -1.503  0.13280
## t_stage4              -0.911657   0.451201 -2.021  0.04333 *
## n_stage2              -0.652606   0.238058 -2.741  0.00612 **
## n_stage3              -0.757283   0.301072 -2.515  0.01189 *
## x6th_stage2            0.069464   0.294083  0.236  0.81327
## x6th_stage3             NA          NA          NA          NA
## x6th_stage4            -0.226947   0.231875 -0.979  0.32771
## x6th_stage5            -0.085418   0.528445 -0.162  0.87159
## differentiate2         0.387813   0.104972  3.694  0.00022 ***
## differentiate3         0.922456   0.193026  4.779  1.76e-06 ***

```

```

## differentiate4      -0.970582   0.533726  -1.819  0.06899 .
## grade2                  NA        NA        NA        NA
## grade3                  NA        NA        NA        NA
## grade4                  NA        NA        NA        NA
## a_stage1                 0.044840  0.266049   0.169  0.86616
## estrogen_status1       0.733142  0.177768   4.124 3.72e-05 ***
## progesterone_status1    0.589919  0.127619   4.623 3.79e-06 ***
## ln_tumor                  -0.053874 0.138931  -0.388  0.69818
## sqrt_examined            0.256137  0.049748   5.149 2.62e-07 ***
## reginol_node_positive   -0.074309  0.015040  -4.941 7.78e-07 ***
## age                      -0.023985  0.005625  -4.264 2.01e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 3444.7 on 4023 degrees of freedom
## Residual deviance: 2951.6 on 3999 degrees of freedom
## AIC: 3001.6
##
## Number of Fisher Scoring iterations: 5

```

Based on the above indicator test summary, we delete grade and x6th\_stage because their output was NA in the output linear model, since NA indicates these predicts may contribute collinearity.

## Model Fitting

### Initial Model

```

glmfit <- glm(status ~ race + marital_status + t_stage + n_stage + differentiate +
                 a_stage + estrogen_status + progesterone_status + ln_tumor +
                 sqrt_examined + reginol_node_positive + age,
                 data = newbc, family = binomial)
summary(glmfit)

##
## Call:
## glm(formula = status ~ race + marital_status + t_stage + n_stage +
##     differentiate + a_stage + estrogen_status + progesterone_status +
##     ln_tumor + sqrt_examined + reginol_node_positive + age, family = binomial,
##     data = newbc)
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)           1.775787  0.587386  3.023  0.002501 **
## race2                -0.522765  0.161744 -3.232  0.001229 **
## race3                 0.419131  0.202312  2.072  0.038293 *
## marital_status2      -0.209325  0.141735 -1.477  0.139709
## marital_status3      -0.140584  0.134542 -1.045  0.296064
## marital_status4      -0.222113  0.192496 -1.154  0.248558
## marital_status5      -0.867369  0.370116 -2.344  0.019104 *

```

```

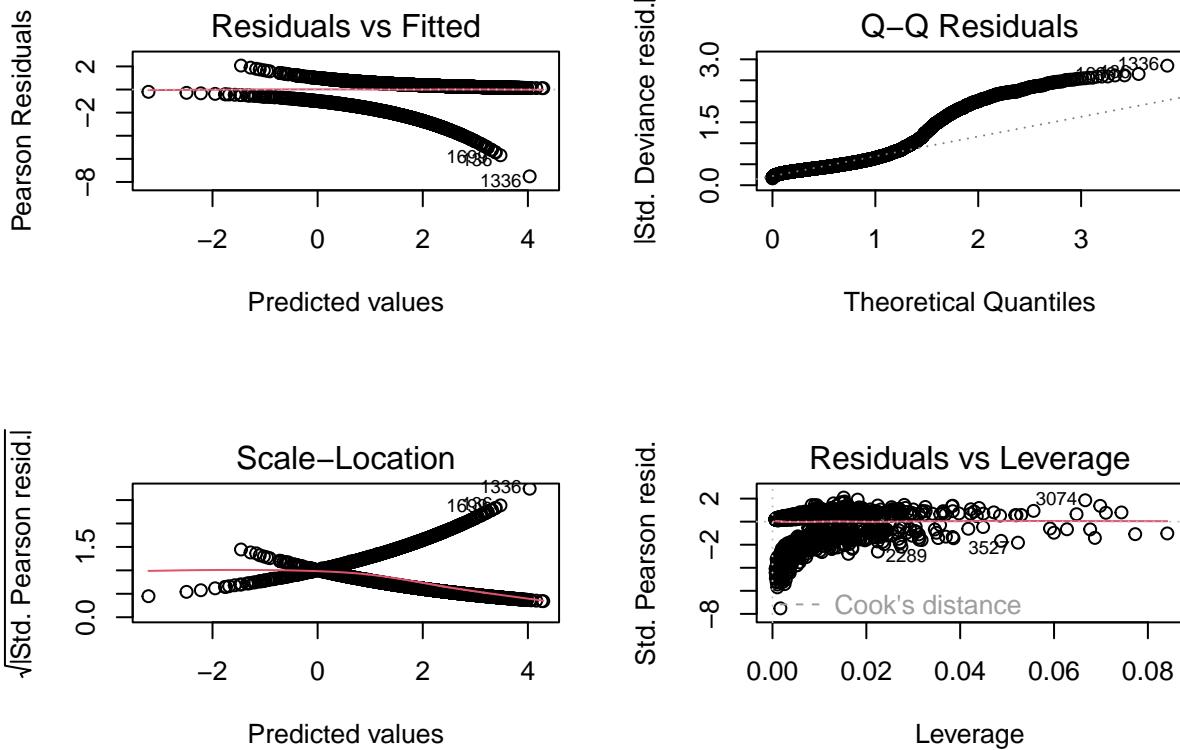
## t_stage2          -0.372488  0.159129 -2.341 0.019243 *
## t_stage3          -0.471905  0.266764 -1.769 0.076894 .
## t_stage4          -1.025962  0.313035 -3.277 0.001047 **
## n_stage2          -0.474091  0.129288 -3.667 0.000245 ***
## n_stage3          -0.637101  0.237639 -2.681 0.007341 **
## differentiate2    0.386027  0.104898  3.680 0.000233 ***
## differentiate3    0.917457  0.192668  4.762 1.92e-06 ***
## differentiate4   -0.947095  0.531127 -1.783 0.074557 .
## a_stage1           0.045211  0.265718  0.170 0.864894
## estrogen_status1  0.737828  0.177538  4.156 3.24e-05 ***
## progesterone_status1 0.588385  0.127499  4.615 3.93e-06 ***
## ln_tumor            -0.055274  0.138631 -0.399 0.690103
## sqrt_examined       0.255997  0.049654  5.156 2.53e-07 ***
## reginol_node_positive -0.074961  0.014984 -5.003 5.65e-07 ***
## age                 -0.023647  0.005618 -4.209 2.56e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 3444.7 on 4023 degrees of freedom
## Residual deviance: 2953.4 on 4002 degrees of freedom
## AIC: 2997.4
##
## Number of Fisher Scoring iterations: 5

```

```

par(mfrow = c(2, 2))
plot(glmfit)

```



## Indicator Variables categorize by Race

```
# According to our data formating in the upper procedure, in Race:  
# 1 represents "white"  
# 2 represents "black"  
# 3 represents "others"  
## A newdataset with "white" as reference  
  
# Ensure race is a factor and not an ordered factor  
bc_ref <- newbc |>  
  mutate(race = factor(race, ordered = FALSE))  
  
# Relevel to make white as reference  
bc_ref <- bc_ref |>  
  mutate(race = relevel(race, ref = 1))  
  
## Run a SLR for race only  
single_race = glm(status ~ race, family = binomial, bc_ref)  
  
summary(single_race)  
  
##  
## Call:  
## glm(formula = status ~ race, family = binomial, data = bc_ref)  
##  
## Coefficients:  
##             Estimate Std. Error z value Pr(>|z|)  
## (Intercept)  1.73909   0.04801 36.221  < 2e-16 ***  
## race2       -0.64505   0.14350 -4.495 6.95e-06 ***  
## race3        0.42389   0.18997  2.231   0.0257 *  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## (Dispersion parameter for binomial family taken to be 1)  
##  
## Null deviance: 3444.7 on 4023 degrees of freedom  
## Residual deviance: 3418.9 on 4021 degrees of freedom  
## AIC: 3424.9  
##  
## Number of Fisher Scoring iterations: 4  
  
## Run a MLR with race ref but without interaction terms  
model_ref = glm(status ~ race + marital_status + t_stage + n_stage + differentiate +  
                 a_stage + estrogen_status + progesterone_status + reginol_node_positive + ln_tumor +  
                 sqrt_examined + age,  
                 data = bc_ref, family = binomial)  
#model_ref  
summary(model_ref)  
  
##  
## Call:  
## glm(formula = status ~ race + marital_status + t_stage + n_stage +
```

```

##      differentiate + a_stage + estrogen_status + progesterone_status +
##      reginol_node_positive + ln_tumor + sqrt_examined + age, family = binomial,
##      data = bc_ref)
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                1.775787  0.587386  3.023 0.002501 **
## race2                     -0.522765  0.161744 -3.232 0.001229 **
## race3                      0.419131  0.202312  2.072 0.038293 *
## marital_status2            -0.209325  0.141735 -1.477 0.139709
## marital_status3            -0.140584  0.134542 -1.045 0.296064
## marital_status4            -0.222113  0.192496 -1.154 0.248558
## marital_status5            -0.867369  0.370116 -2.344 0.019104 *
## t_stage2                   -0.372488  0.159129 -2.341 0.019243 *
## t_stage3                   -0.471905  0.266764 -1.769 0.076894 .
## t_stage4                   -1.025962  0.313035 -3.277 0.001047 **
## n_stage2                   -0.474091  0.129288 -3.667 0.000245 ***
## n_stage3                   -0.637101  0.237639 -2.681 0.007341 **
## differentiate2              0.386027  0.104898  3.680 0.000233 ***
## differentiate3              0.917457  0.192668  4.762 1.92e-06 ***
## differentiate4              -0.947095  0.531127 -1.783 0.074557 .
## a_stage1                   0.045211  0.265718  0.170 0.864894
## estrogen_status1           0.737828  0.177538  4.156 3.24e-05 ***
## progesterone_status1        0.588385  0.127499  4.615 3.93e-06 ***
## reginol_node_positive       -0.074961  0.014984 -5.003 5.65e-07 ***
## ln_tumor                     -0.055274  0.138631 -0.399 0.690103
## sqrt_examined               0.255997  0.049654  5.156 2.53e-07 ***
## age                          -0.023647  0.005618 -4.209 2.56e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 3444.7  on 4023  degrees of freedom
## Residual deviance: 2953.4  on 4002  degrees of freedom
## AIC: 2997.4
##
## Number of Fisher Scoring iterations: 5

## Run a logistic linear with race ref and with interaction terms
model_refinter = glm(status ~ race + marital_status + t_stage + n_stage + differentiate +
                      a_stage + estrogen_status + progesterone_status + reginol_node_positive + ln_tumor +
                      sqrt_examined + age +
                      race*(marital_status + t_stage + n_stage + differentiate +
                            a_stage + estrogen_status + progesterone_status + reginol_node_positive + ln_tumor +
                            sqrt_examined + age),
                      data = bc_ref, family = binomial)
#model_refinter
summary(model_refinter)

##
## Call:
## glm(formula = status ~ race + marital_status + t_stage + n_stage +
##     differentiate + a_stage + estrogen_status + progesterone_status +

```

```

##      reginol_node_positive + ln_tumor + sqrt_examined + age +
##      race * (marital_status + t_stage + n_stage + differentiate +
##                  a_stage + estrogen_status + progesterone_status + reginol_node_positive +
##                  ln_tumor + sqrt_examined + age), family = binomial, data = bc_ref)
##
## Coefficients: (1 not defined because of singularities)
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 2.143663  0.663026  3.233 0.001224 **
## race2                      -3.268199  1.771947 -1.844 0.065123 .
## race3                      5.514563  3.122929  1.766 0.077424 .
## marital_status2            -0.183704  0.152054 -1.208 0.226988
## marital_status3            0.085036  0.158426  0.537 0.591437
## marital_status4            -0.055626  0.220307 -0.252 0.800660
## marital_status5            -0.799662  0.435399 -1.837 0.066266 .
## t_stage2                   -0.275236  0.176881 -1.556 0.119696
## t_stage3                   -0.295508  0.299330 -0.987 0.323529
## t_stage4                   -0.940666  0.342650 -2.745 0.006046 **
## n_stage2                   -0.567320  0.141733 -4.003 6.26e-05 ***
## n_stage3                   -0.843987  0.258302 -3.267 0.001085 **
## differentiate2             0.436042  0.115677  3.769 0.000164 ***
## differentiate3             1.011119  0.214291  4.718 2.38e-06 ***
## differentiate4             -0.948202  0.577004 -1.643 0.100317
## a_stage1                   -0.060677  0.293299 -0.207 0.836106
## estrogen_status1           0.806235  0.196393  4.105 4.04e-05 ***
## progesterone_status1       0.660775  0.138367  4.776 1.79e-06 ***
## reginol_node_positive      -0.065841  0.016154 -4.076 4.59e-05 ***
## ln_tumor                    -0.165689  0.159061 -1.042 0.297566
## sqrt_examined              0.278198  0.054658  5.090 3.58e-07 ***
## age                         -0.027956  0.006220 -4.495 6.97e-06 ***
## race2:marital_status2     -0.231457  0.537505 -0.431 0.666750
## race3:marital_status2     -0.493298  0.712072 -0.693 0.488458
## race2:marital_status3     -1.145440  0.409405 -2.798 0.005145 **
## race3:marital_status3     -1.209396  0.612332 -1.975 0.048261 *
## race2:marital_status4     -0.440236  0.601871 -0.731 0.464507
## race3:marital_status4     -2.132622  0.838528 -2.543 0.010981 *
## race2:marital_status5     0.098692  1.175316  0.084 0.933080
## race3:marital_status5     -0.880739  1.347639 -0.654 0.513407
## race2:t_stage2            -0.465403  0.502263 -0.927 0.354128
## race3:t_stage2            0.437497  0.782923  0.559 0.576298
## race2:t_stage3            -0.703756  0.806589 -0.873 0.382931
## race3:t_stage3            1.603080  1.333557  1.202 0.229322
## race2:t_stage4            0.014416  1.193986  0.012 0.990367
## race3:t_stage4            -0.127448  1.726137 -0.074 0.941142
## race2:n_stage2             0.484228  0.476679  1.016 0.309707
## race3:n_stage2             1.082513  0.645862  1.676 0.093724 .
## race2:n_stage3             1.463405  0.897850  1.630 0.103123
## race3:n_stage3             2.525936  1.462710  1.727 0.084188 .
## race2:differentiate2      -0.208584  0.358607 -0.582 0.560802
## race3:differentiate2      -0.109563  0.465485 -0.235 0.813919
## race2:differentiate3      -0.381274  0.682640 -0.559 0.576483
## race3:differentiate3      -0.140078  0.835274 -0.168 0.866816
## race2:differentiate4      -0.724327  1.669867 -0.434 0.664460
## race3:differentiate4          NA          NA          NA          NA
## race2:a_stage1             1.903928  1.021605  1.864 0.062369 .

```

```

## race3:a_stage1          -1.581496  1.588021 -0.996 0.319303
## race2:estrogen_status1 -0.686042  0.595097 -1.153 0.248982
## race3:estrogen_status1  1.820519  0.981128  1.856 0.063520 .
## race2:progesterone_status1 -0.050826  0.451620 -0.113 0.910394
## race3:progesterone_status1 -2.784800  0.946246 -2.943 0.003251 **
## race2:reginol_node_positive -0.121505  0.064782 -1.876 0.060712 .
## race3:reginol_node_positive -0.153094  0.089011 -1.720 0.085442 .
## race2:ln_tumor            0.776264  0.361491  2.147 0.031762 *
## race3:ln_tumor            -1.194039  0.762242 -1.566 0.117236
## race2:sqrt_examined      -0.085330  0.168637 -0.506 0.612859
## race3:sqrt_examined      -0.162959  0.213679 -0.763 0.445681
## race2:age                 0.008341  0.019360  0.431 0.666599
## race3:age                 0.035269  0.024348  1.449 0.147466
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 3444.7 on 4023 degrees of freedom
## Residual deviance: 2898.0 on 3965 degrees of freedom
## AIC: 3016
##
## Number of Fisher Scoring iterations: 6

## Run a MLR with race ref and with interaction terms selected
model_inter = glm(status ~ race + marital_status + t_stage + n_stage + differentiate +
                   a_stage + estrogen_status + progesterone_status + reginol_node_positive + ln_tumor +
                   sqrt_examined + age +
                   race*(marital_status + progesterone_status + ln_tumor),
                   data = bc_ref, family = binomial)

#model_inter
summary(model_inter)

## 
## Call:
## glm(formula = status ~ race + marital_status + t_stage + n_stage +
##     differentiate + a_stage + estrogen_status + progesterone_status +
##     reginol_node_positive + ln_tumor + sqrt_examined + age +
##     race * (marital_status + progesterone_status + ln_tumor),
##     family = binomial, data = bc_ref)
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                1.882124  0.611049  3.080 0.002069 **
## race2                     -1.131478  0.754804 -1.499 0.133864
## race3                      3.214161  1.374949  2.338 0.019405 *
## marital_status2           -0.186741  0.151851 -1.230 0.218785
## marital_status3           0.086222  0.158014  0.546 0.585299
## marital_status4           -0.076745  0.219242 -0.350 0.726303
## marital_status5           -0.780653  0.435332 -1.793 0.072935 .
## t_stage2                  -0.338441  0.160266 -2.112 0.034708 *
## t_stage3                  -0.375302  0.269563 -1.392 0.163844
## t_stage4                  -0.973119  0.315838 -3.081 0.002063 **
## n_stage2                  -0.495290  0.130427 -3.797 0.000146 ***

```

```

## n_stage3          -0.664599  0.239253 -2.778 0.005473 **
## differentiate2    0.405297  0.105478  3.842 0.000122 ***
## differentiate3    0.947648  0.194613  4.869 1.12e-06 ***
## differentiate4   -0.997840  0.537651 -1.856 0.063464 .
## a_stage1          0.057824  0.268010  0.216 0.829180
## estrogen_status1  0.768781  0.180131  4.268 1.97e-05 ***
## progesterone_status1 0.684689  0.134874  5.077 3.84e-07 ***
## reginol_node_positive -0.075918  0.015092 -5.030 4.89e-07 ***
## ln_tumor           -0.125372  0.146813 -0.854 0.393129
## sqrt_examined     0.258396  0.049886  5.180 2.22e-07 ***
## age                -0.024851  0.005663 -4.389 1.14e-05 ***
## race2:marital_status2 -0.128238  0.518941 -0.247 0.804820
## race3:marital_status2 -0.373279  0.702629 -0.531 0.595238
## race2:marital_status3 -1.109358  0.385796 -2.876 0.004034 **
## race3:marital_status3 -1.349766  0.576548 -2.341 0.019226 *
## race2:marital_status4 -0.427490  0.551410 -0.775 0.438182
## race3:marital_status4 -1.600358  0.671688 -2.383 0.017191 *
## race2:marital_status5 -0.542786  1.006774 -0.539 0.589795
## race3:marital_status5 -0.911159  1.284113 -0.710 0.477975
## race2:progesterone_status1 -0.516652  0.357042 -1.447 0.147887
## race3:progesterone_status1 -1.646000  0.561267 -2.933 0.003361 **
## race2:ln_tumor        0.436542  0.213238  2.047 0.040638 *
## race3:ln_tumor       -0.336268  0.352855 -0.953 0.340594
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 3444.7 on 4023 degrees of freedom
## Residual deviance: 2922.3 on 3990 degrees of freedom
## AIC: 2990.3
##
## Number of Fisher Scoring iterations: 5

```

- From the SRL for single race variable, we can see that the race group of black has a P-value of  $6.95 \times 10^{-6}$  and other race group has a P-value of 0.0257 both smaller than 0.05, so there are statistically significant difference between that to the reference of group white.
- From the MRL model without interaction terms, we can see that the AIC is around 2997.4 which is smaller than the original full model, when for both race2 and race3 we have a P-value smaller than 0.05 indicating significance, so we considering adding interaction terms.
- From the MRL model with race interaction terms added, we can see that the AIC is around 3016 which is getting larger, so we may have too many unnecessary interaction terms, and from the P-values we can see most the interaction covariates are insignificant except some terms in: race:marital\_status, race:progesterone\_status, race:ln\_tumor.
- Therefore, we have the 3rd MRL model with race interaction terms selected, and we get a model with AIC of 2990.3 dropped sharply and the interaction term is included.

## Step-wise: both direction/AIC

### Step-wise for MLR without interaction term:

```

set.seed(123)
# Full Model
full_model <- glm(status ~ race + marital_status + t_stage + n_stage + differentiate +
                     a_stage + estrogen_status + progesterone_status + reginol_node_positive +
                     ln_tumor + sqrt_examined + age,
                     data = newbc, family = binomial)

# Minimal Model (Intercept Only)
min_model <- glm(status ~ 1, data = newbc, family = binomial)

# Forward Selection
forward_model <- step(min_model, scope = list(lower = min_model, upper = full_model),
                      direction = "forward", trace = FALSE)

# Backward Elimination
backward_model <- step(full_model, direction = "backward", trace = FALSE)

# Both Directions
stepwise_model <- step(min_model, scope = list(lower = min_model, upper = full_model),
                       direction = "both", trace = FALSE)

# Print the summary of the chosen model
summary(forward_model)

```

```

##
## Call:
## glm(formula = status ~ n_stage + progesterone_status + differentiate +
##      t_stage + age + race + estrogen_status + sqrt_examined +
##      reginol_node_positive, family = binomial, data = newbc)
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)           1.623139   0.371232   4.372 1.23e-05 ***
## n_stage2            -0.474258   0.128697  -3.685 0.000229 ***
## n_stage3            -0.644148   0.234160  -2.751 0.005943 **
## progesterone_status1 0.600531   0.127305   4.717 2.39e-06 ***
## differentiate2       0.389412   0.104623   3.722 0.000198 ***
## differentiate3       0.908192   0.191600   4.740 2.14e-06 ***
## differentiate4      -0.954165   0.525800  -1.815 0.069571 .
## t_stage2            -0.429230   0.112954  -3.800 0.000145 ***
## t_stage3            -0.561433   0.148563  -3.779 0.000157 ***
## t_stage4            -1.122611   0.242936  -4.621 3.82e-06 ***
## age                 -0.024052   0.005443  -4.419 9.90e-06 ***
## race2              -0.574227   0.158413  -3.625 0.000289 ***
## race3               0.430601   0.201848   2.133 0.032901 *
## estrogen_status1    0.731594   0.176812   4.138 3.51e-05 ***
## sqrt_examined        0.259110   0.049645   5.219 1.80e-07 ***
## reginol_node_positive -0.076301  0.014939  -5.107 3.27e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##

```

```

##      Null deviance: 3444.7  on 4023  degrees of freedom
## Residual deviance: 2961.5  on 4008  degrees of freedom
## AIC: 2993.5
##
## Number of Fisher Scoring iterations: 5

# or summary(backward_model)
summary(backward_model)

## 
## Call:
## glm(formula = status ~ race + t_stage + n_stage + differentiate +
##      estrogen_status + progesterone_status + reginol_node_positive +
##      sqrt_examined + age, family = binomial, data = newbc)
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)           1.623139   0.371232  4.372 1.23e-05 ***
## race2                -0.574227   0.158413 -3.625 0.000289 ***
## race3                 0.430601   0.201848  2.133 0.032901 *
## t_stage2              -0.429230   0.112954 -3.800 0.000145 ***
## t_stage3              -0.561433   0.148563 -3.779 0.000157 ***
## t_stage4              -1.122611   0.242936 -4.621 3.82e-06 ***
## n_stage2              -0.474258   0.128697 -3.685 0.000229 ***
## n_stage3              -0.644148   0.234160 -2.751 0.005943 **
## differentiate2        0.389412   0.104623  3.722 0.000198 ***
## differentiate3        0.908192   0.191600  4.740 2.14e-06 ***
## differentiate4        -0.954165   0.525800 -1.815 0.069571 .
## estrogen_status1     0.731594   0.176812  4.138 3.51e-05 ***
## progesterone_status1 0.600531   0.127305  4.717 2.39e-06 ***
## reginol_node_positive -0.076301   0.014939 -5.107 3.27e-07 ***
## sqrt_examined         0.259110   0.049645  5.219 1.80e-07 ***
## age                  -0.024052   0.005443 -4.419 9.90e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 3444.7  on 4023  degrees of freedom
## Residual deviance: 2961.5  on 4008  degrees of freedom
## AIC: 2993.5
##
## Number of Fisher Scoring iterations: 5

```

```

# or summary(stepwise_model)
summary(stepwise_model)

```

```

## 
## Call:
## glm(formula = status ~ n_stage + progesterone_status + differentiate +
##      t_stage + age + race + estrogen_status + sqrt_examined +
##      reginol_node_positive, family = binomial, data = newbc)
## 
```

```

## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)           1.623139   0.371232  4.372 1.23e-05 ***
## n_stage2            -0.474258   0.128697 -3.685 0.000229 ***
## n_stage3            -0.644148   0.234160 -2.751 0.005943 **
## progesterone_status1 0.600531   0.127305  4.717 2.39e-06 ***
## differentiate2       0.389412   0.104623  3.722 0.000198 ***
## differentiate3       0.908192   0.191600  4.740 2.14e-06 ***
## differentiate4      -0.954165   0.525800 -1.815 0.069571 .
## t_stage2             -0.429230   0.112954 -3.800 0.000145 ***
## t_stage3             -0.561433   0.148563 -3.779 0.000157 ***
## t_stage4             -1.122611   0.242936 -4.621 3.82e-06 ***
## age                  -0.024052   0.005443 -4.419 9.90e-06 ***
## race2                -0.574227   0.158413 -3.625 0.000289 ***
## race3                0.430601   0.201848  2.133 0.032901 *
## estrogen_status1     0.731594   0.176812  4.138 3.51e-05 ***
## sqrt_examined        0.259110   0.049645  5.219 1.80e-07 ***
## reginol_node_positive -0.076301   0.014939 -5.107 3.27e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 3444.7 on 4023 degrees of freedom
## Residual deviance: 2961.5 on 4008 degrees of freedom
## AIC: 2993.5
##
## Number of Fisher Scoring iterations: 5

```

From the both direction step-wise, the best model without interaction term has lowest AIC of 2993.5 is:  
 $\text{status} \sim \text{n\_stage} + \text{progesterone\_status} + \text{differentiate} + \text{t\_stage} + \text{age} + \text{race} + \text{estrogen\_status} + \text{sqrt\_examined} + \text{reginol\_node\_positive}$

#### Step-wise for MLR with interaction term:

```

# Full Model with Interaction Terms
full_interaction_model <- glm(status ~ race + marital_status + t_stage + n_stage + differentiate +
                                a_stage + estrogen_status + progesterone_status + reginol_node_positive + ln_tumor +
                                sqrt_examined + age +
                                race*(marital_status + progesterone_status + ln_tumor),
                                data = bc_ref, family = binomial)

# Minimal Model (Intercept Only)
min_model <- glm(status ~ 1, data = bc_ref, family = binomial)

# Stepwise Selection (Forward, Backward, or Both)
stepwise_interaction_model <- step(min_model, scope = list(lower = min_model, upper = full_interaction_model),
                                     direction = "both", trace = FALSE)

# Print the summary of the chosen model
summary(stepwise_interaction_model)

```

```

## 
## Call:
## glm(formula = status ~ n_stage + progesterone_status + differentiate +
##      t_stage + age + race + estrogen_status + sqrt_examined +
##      reginol_node_positive + marital_status + progesterone_status:race +
##      race:marital_status, family = binomial, data = bc_ref)
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 1.570840  0.383329  4.098 4.17e-05 ***
## n_stage2                  -0.496361  0.129955 -3.819 0.000134 ***
## n_stage3                  -0.679178  0.236559 -2.871 0.004091 **
## progesterone_status1       0.687634  0.134627  5.108 3.26e-07 ***
## differentiate2              0.401059  0.105340  3.807 0.000140 ***
## differentiate3              0.952710  0.194356  4.902 9.49e-07 ***
## differentiate4             -0.947664  0.540209 -1.754 0.079387 .
## t_stage2                  -0.406468  0.113944 -3.567 0.000361 ***
## t_stage3                  -0.534698  0.149840 -3.568 0.000359 ***
## t_stage4                  -1.100569  0.245075 -4.491 7.10e-06 ***
## age                        -0.024541  0.005655 -4.340 1.43e-05 ***
## race2                      0.224318  0.383613  0.585 0.558716
## race3                      2.011663  0.542750  3.706 0.000210 ***
## estrogen_status1            0.769528  0.179823  4.279 1.87e-05 ***
## sqrt_examined                0.257377  0.049803  5.168 2.37e-07 ***
## reginol_node_positive       -0.075272  0.015097 -4.986 6.16e-07 ***
## marital_status2             -0.191812  0.151472 -1.266 0.205399
## marital_status3             0.078126  0.157546  0.496 0.619971
## marital_status4             -0.082890  0.218859 -0.379 0.704885
## marital_status5             -0.779425  0.434517 -1.794 0.072849 .
## progesterone_status1:race2 -0.465115  0.359360 -1.294 0.195566
## progesterone_status1:race3 -1.605449  0.549534 -2.921 0.003484 **
## race2:marital_status2      -0.036425  0.525651 -0.069 0.944755
## race3:marital_status2      -0.293014  0.695570 -0.421 0.673567
## race2:marital_status3      -1.051902  0.387355 -2.716 0.006616 **
## race3:marital_status3      -1.315085  0.567928 -2.316 0.020581 *
## race2:marital_status4      -0.421867  0.561169 -0.752 0.452193
## race3:marital_status4      -1.585759  0.660762 -2.400 0.016400 *
## race2:marital_status5      -0.519846  1.015055 -0.512 0.608556
## race3:marital_status5      -0.931811  1.284515 -0.725 0.468195
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 3444.7 on 4023 degrees of freedom
## Residual deviance: 2927.9 on 3994 degrees of freedom
## AIC: 2987.9
##
## Number of Fisher Scoring iterations: 5

```

From the both direction step-wise, the best model with interaction term has lowest AIC of 2987.9 is: status ~ n\_stage + progesterone\_status + differentiate + t\_stage + age + race + estrogen\_status + sqrt\_examined + reginol\_node\_positive + marital\_status + progesterone\_status:race + race:marital\_status.

## VIF

```
# Calculate VIF for the model without interaction
vif_result_bic <- vif(stepwise_model)
print(vif_result_bic)

##                                     GVIF Df GVIF^(1/(2*Df))
## n_stage                  3.835937  2     1.399484
## progesterone_status      1.426727  1     1.194457
## differentiate            1.112306  3     1.017897
## t_stage                  1.098462  3     1.015775
## age                      1.042599  1     1.021077
## race                      1.020862  2     1.005175
## estrogen_status           1.471867  1     1.213205
## sqrt_examined             1.423764  1     1.193216
## reginol_node_positive    4.073707  1     2.018343

# Calculate VIF for the interaction model
vif_result_aic <- vif(stepwise_interaction_model, type = 'predictor')

## Warning in vif.lm(stepwise_interaction_model, type = "predictor"): type = 'predictor' is available or
##   type = 'terms' will be used

print(vif_result_aic)

##                                     GVIF Df GVIF^(1/(2*Df))
## n_stage                  3.884443  2     1.403887
## progesterone_status      1.565824  1     1.251329
## differentiate            1.133420  3     1.021093
## t_stage                  1.114696  3     1.018262
## age                      1.120454  1     1.058515
## race                      43.194842  2     2.563646
## estrogen_status           1.476767  1     1.215223
## sqrt_examined             1.426509  1     1.194365
## reginol_node_positive    4.094125  1     2.023394
## marital_status            3.129525  4     1.153280
## progesterone_status:race 24.566818  2     2.226318
## race:marital_status       14.850111  8     1.183677
```

## Partial Test

### Partial Test for binary Y

```
# Model without Interaction Terms
model_no_interaction <- glm(status ~ n_stage + progesterone_status + differentiate + t_stage +
                             age + race + estrogen_status + sqrt_examined + reginol_node_positive,
                             data = newbc, family = binomial)

# Model with Selected Interaction Terms
```

```

model_with_interaction <- glm(status ~ n_stage + progesterone_status + differentiate + t_stage +
                               age + race + estrogen_status + sqrt_examined + reginol_node_positive +
                               marital_status + progesterone_status:race + race:marital_status,
                               data = bc_ref, family = binomial)

# Partial Test
anova(model_no_interaction, model_with_interaction, test = "Chisq")

```

```

## Analysis of Deviance Table
##
## Model 1: status ~ n_stage + progesterone_status + differentiate + t_stage +
##           age + race + estrogen_status + sqrt_examined + reginol_node_positive
## Model 2: status ~ n_stage + progesterone_status + differentiate + t_stage +
##           age + race + estrogen_status + sqrt_examined + reginol_node_positive +
##           marital_status + progesterone_status:race + race:marital_status
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      4008    2961.5
## 2      3994    2927.9 14    33.565 0.002386 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

The p-value is below the common alpha level of 0.05, indicating that the inclusion of interaction terms in the model significantly improves the fit compared to the model without interaction terms. This result supports selecting the model with interaction terms, as it provides a better explanation of the variation in the response variable (status).

## ANOVA

```

anova_result1 <- anova(model_with_interaction)
print(anova_result1)

```

anova for binary Y

```

## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: status
##
## Terms added sequentially (first to last)
##
##
##                                     Df Deviance Resid. Df Resid. Dev
## NULL                               4023    3444.7
## n_stage                            2   230.650   4021    3214.0
## progesterone_status                1    85.373   4020    3128.7
## differentiate                       3    43.629   4017    3085.0
## t_stage                            3    27.748   4014    3057.3
## age                                1    19.891   4013    3037.4

```

```

## race              2   18.417    4011   3019.0
## estrogen_status  1   15.880    4010   3003.1
## sqrt_examined   1   15.282    4009   2987.8
## reginol_node_positive 1   26.334    4008   2961.5
## marital_status   4   7.907     4004   2953.6
## progesterone_status:race 2   9.500     4002   2944.1
## race:marital_status 8   16.158    3994   2927.9

summary(model_with_interaction)

## 
## Call:
## glm(formula = status ~ n_stage + progesterone_status + differentiate +
##      t_stage + age + race + estrogen_status + sqrt_examined +
##      reginol_node_positive + marital_status + progesterone_status:race +
##      race:marital_status, family = binomial, data = bc_ref)
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                1.570840  0.383329  4.098 4.17e-05 ***
## n_stage2                  -0.496361  0.129955 -3.819 0.000134 ***
## n_stage3                  -0.679178  0.236559 -2.871 0.004091 **
## progesterone_status1       0.687634  0.134627  5.108 3.26e-07 ***
## differentiate2             0.401059  0.105340  3.807 0.000140 ***
## differentiate3             0.952710  0.194356  4.902 9.49e-07 ***
## differentiate4             -0.947664  0.540209 -1.754 0.079387 .
## t_stage2                  -0.406468  0.113944 -3.567 0.000361 ***
## t_stage3                  -0.534698  0.149840 -3.568 0.000359 ***
## t_stage4                  -1.100569  0.245075 -4.491 7.10e-06 ***
## age                        -0.024541  0.005655 -4.340 1.43e-05 ***
## race2                      0.224318  0.383613  0.585 0.558716
## race3                      2.011663  0.542750  3.706 0.000210 ***
## estrogen_status1            0.769528  0.179823  4.279 1.87e-05 ***
## sqrt_examined              0.257377  0.049803  5.168 2.37e-07 ***
## reginol_node_positive      -0.075272  0.015097 -4.986 6.16e-07 ***
## marital_status2            -0.191812  0.151472 -1.266 0.205399
## marital_status3            0.078126  0.157546  0.496 0.619971
## marital_status4            -0.082890  0.218859 -0.379 0.704885
## marital_status5            -0.779425  0.434517 -1.794 0.072849 .
## progesterone_status1:race2 -0.465115  0.359360 -1.294 0.195566
## progesterone_status1:race3 -1.605449  0.549534 -2.921 0.003484 **
## race2:marital_status2      -0.036425  0.525651 -0.069 0.944755
## race3:marital_status2      -0.293014  0.695570 -0.421 0.673567
## race2:marital_status3      -1.051902  0.387355 -2.716 0.006616 **
## race3:marital_status3      -1.315085  0.567928 -2.316 0.020581 *
## race2:marital_status4      -0.421867  0.561169 -0.752 0.452193
## race3:marital_status4      -1.585759  0.660762 -2.400 0.016400 *
## race2:marital_status5      -0.519846  1.015055 -0.512 0.608556
## race3:marital_status5      -0.931811  1.284515 -0.725 0.468195
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##

```

```

##      Null deviance: 3444.7  on 4023  degrees of freedom
## Residual deviance: 2927.9  on 3994  degrees of freedom
## AIC: 2987.9
##
## Number of Fisher Scoring iterations: 5

null_model <- glm(status ~ 1, data = bc_ref, family = binomial)

anova(null_model, model_with_interaction, test = "Chisq")

## Analysis of Deviance Table
##
## Model 1: status ~ 1
## Model 2: status ~ n_stage + progesterone_status + differentiate + t_stage +
##           age + race + estrogen_status + sqrt_examined + reginol_node_positive +
##           marital_status + progesterone_status:race + race:marital_status
##           Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      4023    3444.7
## 2      3994    2927.9 29     516.77 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

With p significant p value, we have enough evidence to show that the full model is better, which is status ~ n\_stage + progesterone\_status + differentiate + t\_stage + age + race + estrogen\_status + sqrt\_examined + reginol\_node\_positive + marital\_status + progesterone\_status:race + race:marital\_status

## Criterion procedures

### Model with Interaction terms

```

aic_selected_bin_model <- stepAIC(model_with_interaction, direction = "both")

## Start:  AIC=2987.91
## status ~ n_stage + progesterone_status + differentiate + t_stage +
##           age + race + estrogen_status + sqrt_examined + reginol_node_positive +
##           marital_status + progesterone_status:race + race:marital_status
##
##                               Df Deviance     AIC
## <none>                      2927.9 2987.9
## - race:marital_status       8   2944.1 2988.1
## - progesterone_status:race  2   2939.1 2995.1
## - n_stage                   2   2943.2 2999.2
## - estrogen_status            1   2946.0 3004.0
## - age                        1   2947.1 3005.1
## - t_stage                    3   2956.0 3010.0
## - reginol_node_positive      1   2953.0 3011.0
## - sqrt_examined              1   2955.0 3013.0
## - differentiate               3   2963.2 3017.2

```

```

summary(aic_selected_bin_model)

##
## Call:
## glm(formula = status ~ n_stage + progesterone_status + differentiate +
##      t_stage + age + race + estrogen_status + sqrt_examined +
##      reginol_node_positive + marital_status + progesterone_status:race +
##      race:marital_status, family = binomial, data = bc_ref)
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                1.570840  0.383329  4.098 4.17e-05 ***
## n_stage2                 -0.496361  0.129955 -3.819 0.000134 ***
## n_stage3                 -0.679178  0.236559 -2.871 0.004091 **
## progesterone_status1      0.687634  0.134627  5.108 3.26e-07 ***
## differentiate2            0.401059  0.105340  3.807 0.000140 ***
## differentiate3            0.952710  0.194356  4.902 9.49e-07 ***
## differentiate4            -0.947664  0.540209 -1.754 0.079387 .
## t_stage2                  -0.406468  0.113944 -3.567 0.000361 ***
## t_stage3                  -0.534698  0.149840 -3.568 0.000359 ***
## t_stage4                  -1.100569  0.245075 -4.491 7.10e-06 ***
## age                        -0.024541  0.005655 -4.340 1.43e-05 ***
## race2                      0.224318  0.383613  0.585 0.558716
## race3                      2.011663  0.542750  3.706 0.000210 ***
## estrogen_status1           0.769528  0.179823  4.279 1.87e-05 ***
## sqrt_examined              0.257377  0.049803  5.168 2.37e-07 ***
## reginol_node_positive      -0.075272  0.015097 -4.986 6.16e-07 ***
## marital_status2            -0.191812  0.151472 -1.266 0.205399
## marital_status3            0.078126  0.157546  0.496 0.619971
## marital_status4            -0.082890  0.218859 -0.379 0.704885
## marital_status5            -0.779425  0.434517 -1.794 0.072849 .
## progesterone_status1:race2 -0.465115  0.359360 -1.294 0.195566
## progesterone_status1:race3 -1.605449  0.549534 -2.921 0.003484 **
## race2:marital_status2      -0.036425  0.525651 -0.069 0.944755
## race3:marital_status2      -0.293014  0.695570 -0.421 0.673567
## race2:marital_status3      -1.051902  0.387355 -2.716 0.006616 **
## race3:marital_status3      -1.315085  0.567928 -2.316 0.020581 *
## race2:marital_status4      -0.421867  0.561169 -0.752 0.452193
## race3:marital_status4      -1.585759  0.660762 -2.400 0.016400 *
## race2:marital_status5      -0.519846  1.015055 -0.512 0.608556
## race3:marital_status5      -0.931811  1.284515 -0.725 0.468195
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 3444.7  on 4023  degrees of freedom
## Residual deviance: 2927.9  on 3994  degrees of freedom
## AIC: 2987.9
##
## Number of Fisher Scoring iterations: 5

```

```

bic_selected_bin_model <- stepAIC(model_with_interaction, direction = "both", family = binomial, k = log2(nrow(data)) + 1)

## Start: AIC=3176.91
## status ~ n_stage + progesterone_status + differentiate + t_stage +
##      age + race + estrogen_status + sqrt_examined + reginol_node_positive +
##      marital_status + progesterone_status:race + race:marital_status
##
##                                     Df Deviance    AIC
## - race:marital_status          8   2944.1 3126.7
## - progesterone_status:race    2   2939.1 3171.5
## - n_stage                      2   2943.2 3175.6
## <none>                         2927.9 3176.9
## - t_stage                      3   2956.0 3180.1
## - estrogen_status              1   2946.0 3186.7
## - differentiate                3   2963.2 3187.3
## - age                          1   2947.1 3187.8
## - reginol_node_positive        1   2953.0 3193.7
## - sqrt_examined                1   2955.0 3195.7
##
## Step: AIC=3126.67
## status ~ n_stage + progesterone_status + differentiate + t_stage +
##      age + race + estrogen_status + sqrt_examined + reginol_node_positive +
##      marital_status + progesterone_status:race
##
##                                     Df Deviance    AIC
## - marital_status               4   2952.3 3101.7
## - progesterone_status:race    2   2953.6 3119.6
## - n_stage                      2   2957.9 3123.9
## <none>                         2944.1 3126.7
## - t_stage                      3   2973.0 3130.7
## - differentiate                3   2977.7 3135.4
## - estrogen_status              1   2962.7 3137.0
## - age                          1   2962.9 3137.2
## - reginol_node_positive        1   2970.2 3144.5
## - sqrt_examined                1   2971.1 3145.4
## + race:marital_status          8   2927.9 3176.9
##
## Step: AIC=3101.69
## status ~ n_stage + progesterone_status + differentiate + t_stage +
##      age + race + estrogen_status + sqrt_examined + reginol_node_positive +
##      progesterone_status:race
##
##                                     Df Deviance    AIC
## - progesterone_status:race    2   2961.5 3094.3
## - n_stage                      2   2966.0 3098.8
## <none>                         2952.3 3101.7
## - t_stage                      3   2982.2 3106.7
## - differentiate                3   2985.8 3110.3
## - estrogen_status              1   2970.7 3111.8
## - age                          1   2973.1 3114.2
## - reginol_node_positive        1   2979.5 3120.6
## - sqrt_examined                1   2980.1 3121.2
## + marital_status                4   2944.1 3126.7

```

```

## 
## Step: AIC=3094.28
## status ~ n_stage + progesterone_status + differentiate + t_stage +
##      age + race + estrogen_status + sqrt_examined + reginol_node_positive
##
##                                     Df Deviance    AIC
## - n_stage                         2   2975.7 3091.9
## <none>                           2961.5 3094.3
## - race                            2   2980.3 3096.5
## - t_stage                          3   2992.1 3100.0
## + progesterone_status:race        2   2952.3 3101.7
## - estrogen_status                 1   2978.4 3102.9
## - differentiate                   3   2995.1 3103.0
## - age                             1   2981.4 3105.9
## - progesterone_status             1   2982.6 3107.1
## - reginol_node_positive           1   2987.8 3112.3
## - sqrt_examined                  1   2989.1 3113.6
## + marital_status                  4   2953.6 3119.6
##
## Step: AIC=3091.86
## status ~ progesterone_status + differentiate + t_stage + age +
##      race + estrogen_status + sqrt_examined + reginol_node_positive
##
##                                     Df Deviance    AIC
## <none>                           2975.7 3091.9
## + n_stage                          2   2961.5 3094.3
## - race                            2   2995.5 3095.1
## + progesterone_status:race        2   2966.0 3098.8
## - estrogen_status                 1   2993.4 3101.3
## - age                             1   2995.6 3103.5
## - differentiate                   3   3012.4 3103.7
## - t_stage                          3   3013.8 3105.1
## - progesterone_status             1   2997.5 3105.4
## - sqrt_examined                  1   3002.0 3109.9
## + marital_status                  4   2967.9 3117.3
## - reginol_node_positive           1   3128.6 3236.5

summary(bic_selected_bin_model)

```

```

## 
## Call:
## glm(formula = status ~ progesterone_status + differentiate +
##      t_stage + age + race + estrogen_status + sqrt_examined +
##      reginol_node_positive, family = binomial, data = bc_ref)
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)               1.597330  0.369903  4.318  1.57e-05 ***
## progesterone_status1     0.610172  0.126822  4.811  1.50e-06 ***
## differentiate2            0.408222  0.104130  3.920  8.84e-05 ***
## differentiate3            0.949277  0.191071  4.968  6.76e-07 ***
## differentiate4           -0.925654  0.516388 -1.793  0.073044 .
## t_stage2                -0.476223  0.112076 -4.249  2.15e-05 ***
## t_stage3                -0.636738  0.147043 -4.330  1.49e-05 ***

```

```

## t_stage4          -1.211832  0.241560  -5.017 5.26e-07 ***
## age              -0.024010  0.005429  -4.422 9.76e-06 ***
## race2            -0.595527  0.157330  -3.785 0.000154 ***
## race3            0.425009  0.201922   2.105 0.035307 *
## estrogen_status1 0.744754  0.175621   4.241 2.23e-05 ***
## sqrt_examined     0.249258  0.048899   5.097 3.44e-07 ***
## reginol_node_positive -0.108237  0.008844 -12.239 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 3444.7 on 4023 degrees of freedom
## Residual deviance: 2975.7 on 4010 degrees of freedom
## AIC: 3003.7
##
## Number of Fisher Scoring iterations: 5

```

In the interaction model, the AIC selected model is status ~ race + marital\_status + t\_stage + n\_stage + differentiate + estrogen\_status + progesterone\_status + reginol\_node\_positive + ln\_tumor + sqrt\_examined + age + race:marital\_status + race:progesterone\_status + race:ln\_tumor.

The BIC selected model is status ~ race + differentiate + estrogen\_status + progesterone\_status + reginol\_node\_positive + ln\_tumor + sqrt\_examined + age.

### Model without Interaction Terms

```
aic_selected_bin_model <- stepAIC(model_no_interaction, direction = "both")
```

```

## Start:  AIC=2993.48
## status ~ n_stage + progesterone_status + differentiate + t_stage +
##       age + race + estrogen_status + sqrt_examined + reginol_node_positive
##
##                               Df Deviance    AIC
## <none>                      2961.5 2993.5
## - n_stage                    2    2975.7 3003.7
## - race                       2    2980.3 3008.3
## - estrogen_status             1    2978.4 3008.4
## - age                        1    2981.4 3011.4
## - progesterone_status         1    2982.6 3012.6
## - reginol_node_positive      1    2987.8 3017.8
## - t_stage                     3    2992.1 3018.1
## - sqrt_examined               1    2989.1 3019.1
## - differentiate                3    2995.1 3021.1

```

```
summary(aic_selected_bin_model)
```

```

##
## Call:
## glm(formula = status ~ n_stage + progesterone_status + differentiate +
##       t_stage + age + race + estrogen_status + sqrt_examined +

```

```

##      reginol_node_positive, family = binomial, data = newbc)
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)           1.623139  0.371232  4.372 1.23e-05 ***
## n_stage2            -0.474258  0.128697 -3.685 0.000229 ***
## n_stage3            -0.644148  0.234160 -2.751 0.005943 **
## progesterone_status1 0.600531  0.127305  4.717 2.39e-06 ***
## differentiate2       0.389412  0.104623  3.722 0.000198 ***
## differentiate3       0.908192  0.191600  4.740 2.14e-06 ***
## differentiate4      -0.954165  0.525800 -1.815 0.069571 .
## t_stage2            -0.429230  0.112954 -3.800 0.000145 ***
## t_stage3            -0.561433  0.148563 -3.779 0.000157 ***
## t_stage4            -1.122611  0.242936 -4.621 3.82e-06 ***
## age                 -0.024052  0.005443 -4.419 9.90e-06 ***
## race2              -0.574227  0.158413 -3.625 0.000289 ***
## race3               0.430601  0.201848  2.133 0.032901 *
## estrogen_status1    0.731594  0.176812  4.138 3.51e-05 ***
## sqrt_examined       0.259110  0.049645  5.219 1.80e-07 ***
## reginol_node_positive -0.076301  0.014939 -5.107 3.27e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 3444.7 on 4023 degrees of freedom
## Residual deviance: 2961.5 on 4008 degrees of freedom
## AIC: 2993.5
##
## Number of Fisher Scoring iterations: 5

bic_selected_bin_model <- stepAIC(model_no_interaction, direction = "both", family = binomial, k = log(2))

## Start:  AIC=3094.28
## status ~ n_stage + progesterone_status + differentiate + t_stage +
##        age + race + estrogen_status + sqrt_examined + reginol_node_positive
##
##                               Df Deviance     AIC
## - n_stage                  2   2975.7 3091.9
## <none>                      2961.5 3094.3
## - race                     2   2980.3 3096.5
## - t_stage                   3   2992.1 3100.0
## - estrogen_status           1   2978.4 3102.9
## - differentiate             3   2995.1 3103.0
## - age                       1   2981.4 3105.9
## - progesterone_status       1   2982.6 3107.1
## - reginol_node_positive     1   2987.8 3112.3
## - sqrt_examined             1   2989.1 3113.6
##
## Step:  AIC=3091.86
## status ~ progesterone_status + differentiate + t_stage + age +
##        race + estrogen_status + sqrt_examined + reginol_node_positive
##
##                               Df Deviance     AIC
```

```

## <none>                      2975.7 3091.9
## + n_stage                     2    2961.5 3094.3
## - race                         2    2995.5 3095.1
## - estrogen_status                1    2993.4 3101.3
## - age                           1    2995.6 3103.5
## - differentiate                 3    3012.4 3103.7
## - t_stage                       3    3013.8 3105.1
## - progesterone_status            1    2997.5 3105.4
## - sqrt_examined                  1    3002.0 3109.9
## - reginol_node_positive          1    3128.6 3236.5

summary(bic_selected_bin_model)

##
## Call:
## glm(formula = status ~ progesterone_status + differentiate +
##      t_stage + age + race + estrogen_status + sqrt_examined +
##      reginol_node_positive, family = binomial, data = newbc)
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)           1.597330   0.369903  4.318 1.57e-05 ***
## progesterone_status1 0.610172   0.126822  4.811 1.50e-06 ***
## differentiate2        0.408222   0.104130  3.920 8.84e-05 ***
## differentiate3        0.949277   0.191071  4.968 6.76e-07 ***
## differentiate4       -0.925654   0.516388 -1.793 0.073044 .
## t_stage2              -0.476223   0.112076 -4.249 2.15e-05 ***
## t_stage3              -0.636738   0.147043 -4.330 1.49e-05 ***
## t_stage4              -1.211832   0.241560 -5.017 5.26e-07 ***
## age                   -0.024010   0.005429 -4.422 9.76e-06 ***
## race2                 -0.595527   0.157330 -3.785 0.000154 ***
## race3                 0.425009   0.201922  2.105 0.035307 *
## estrogen_status1      0.744754   0.175621  4.241 2.23e-05 ***
## sqrt_examined         0.249258   0.048899  5.097 3.44e-07 ***
## reginol_node_positive -0.108237   0.008844 -12.239 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 3444.7 on 4023 degrees of freedom
## Residual deviance: 2975.7 on 4010 degrees of freedom
## AIC: 3003.7
##
## Number of Fisher Scoring iterations: 5

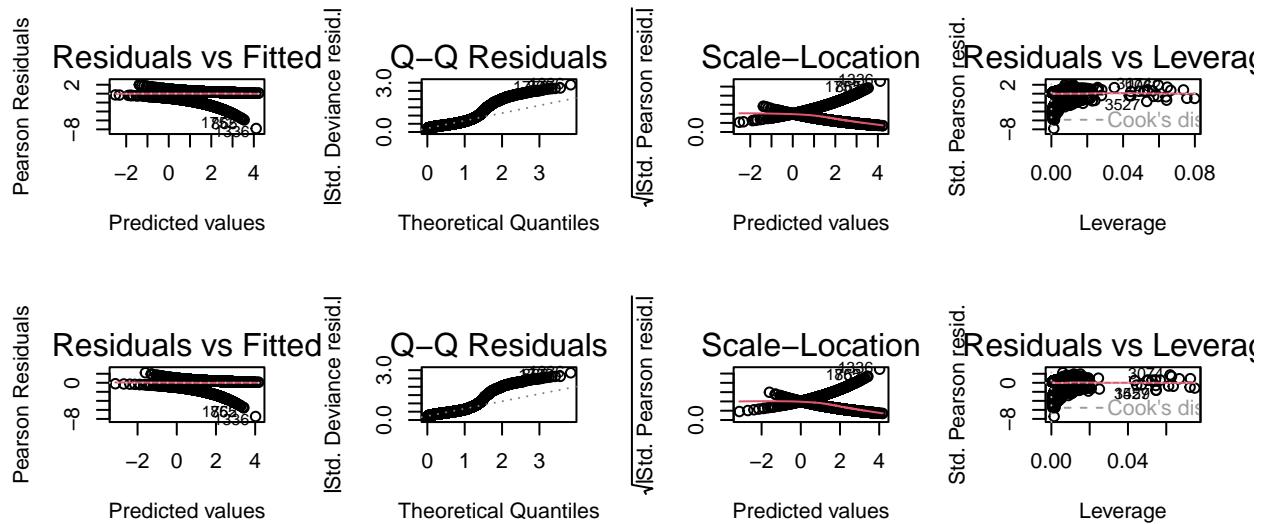
```

In the model without interaction, the AIC selected model is  $\text{status} \sim \text{n\_stage} + \text{progesterone\_status} + \text{differentiate} + \text{t\_stage} + \text{age} + \text{race} + \text{estrogen\_status} + \text{sqrt\_examined} + \text{reginol\_node\_positive}$ , having a AIC value of 2961.5.

The BIC selected model is  $\text{status} \sim \text{progesterone\_status} + \text{differentiate} + \text{t\_stage} + \text{age} + \text{race} + \text{estrogen\_status} + \text{sqrt\_examined} + \text{reginol\_node\_positive}$  with AIC value of 3003.7.

Plots to check the model assumptions:

```
par(mfrow = c(3,4))
plot(aic_selected_bin_model)
plot(bic_selected_bin_model)
```



3.0 transformation edited 3.1 interaction transformation ?? 3.2 partial test 3.3 diagnostic boxcox 4. Stepwise: forward/ backward /AIC 5. final model 6. model assumption (check multicollinearity (VIF)) 7. cross validation