

Data analysis

12/09/2023

Libraries

```
library(tidyverse)
library(readr)
library(boot)
library(table1)
library(gridExtra)
library(MASS)
library(car)
library(leaps)
library(corrplot)
library(caret)
library(car)
library(ResourceSelection)
library(pROC)
```

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

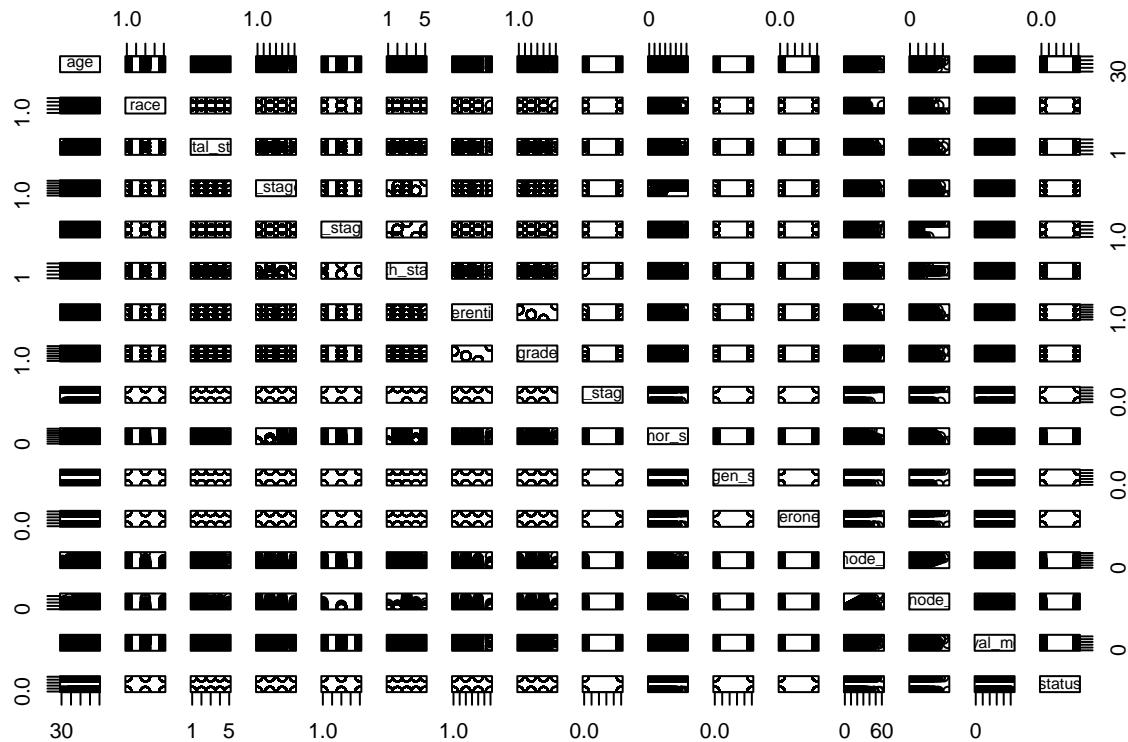
```

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")

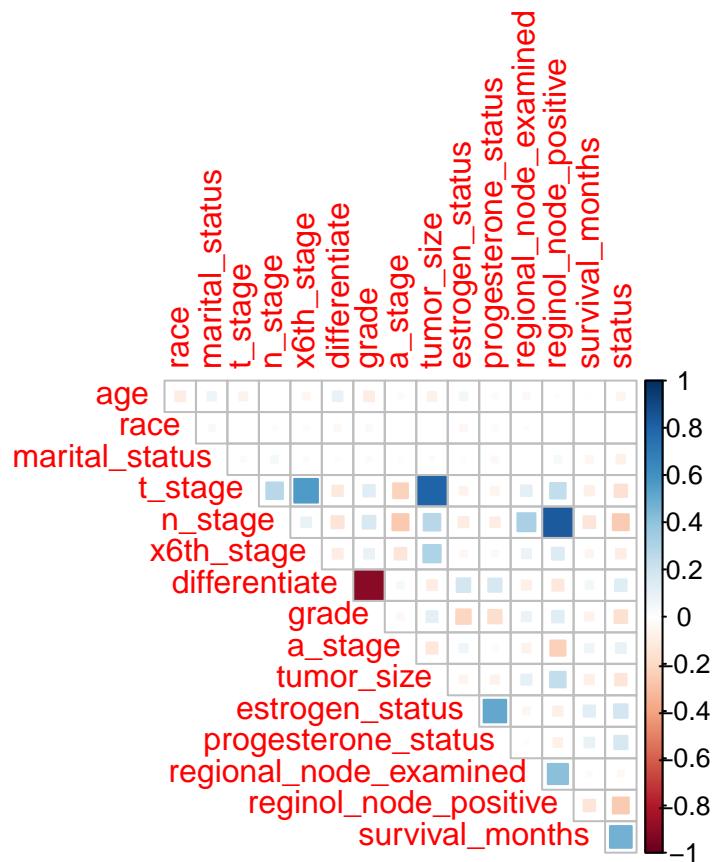
```

Table 1: Correlation for all variables

| | age | race | marital | <u>stage</u> | <u>stage6</u> | <u>differentiate</u> | <u>grade</u> | <u>stage</u> | <u>noes</u> | <u>size</u> | <u>progesterone</u> | <u>status</u> | <u>regional</u> | <u>node</u> | <u>survival</u> | <u>months</u> |
|---------------------|-------------------|-------------------|---------|--------------|---------------|----------------------|--------------|--------------|-------------|-------------|---------------------|---------------|-----------------|-------------|-----------------|---------------|
| age | 1.0000- | 0.0755 | - | 0.0029 | - | 0.0932 | - | 0.0209 | - | 0.0598 | - | - | - | 0.0126 | - | - |
| | | 0.0970 | | 0.0669 | 0.0450 | | 0.0993 | | 0.0772 | | 0.0213 | 0.0333 | | | 0.0094 | 0.0559 |
| 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 | 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 | 0.0170 | 0.0221 | 0.5630 | 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 | 0.1100 | 0.1488 | 0.0999 | 0.9083 | 0.0995 | 0.0834 | 0.1229 | |
| grade | - | 0.0300 | 0.0206 | 0.1310 | 0.1620 | 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 | 0.2600 | 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 | 0.0598 | - | - | - | - | 0.1868 | - | 0.0656 | - | 1.0000 | 0.5133 | - | - | 0.1285 | 0.1847 | |
| | | | | | | | 0.0420 | 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 | 0.0181 | 0.0781 | |
| regional_node | 0.0110 | 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_survival | 0.0010 | 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
```

```
plot7differentiate =  
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"  
)
```

```
#plot7differentiate
```

```
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

```

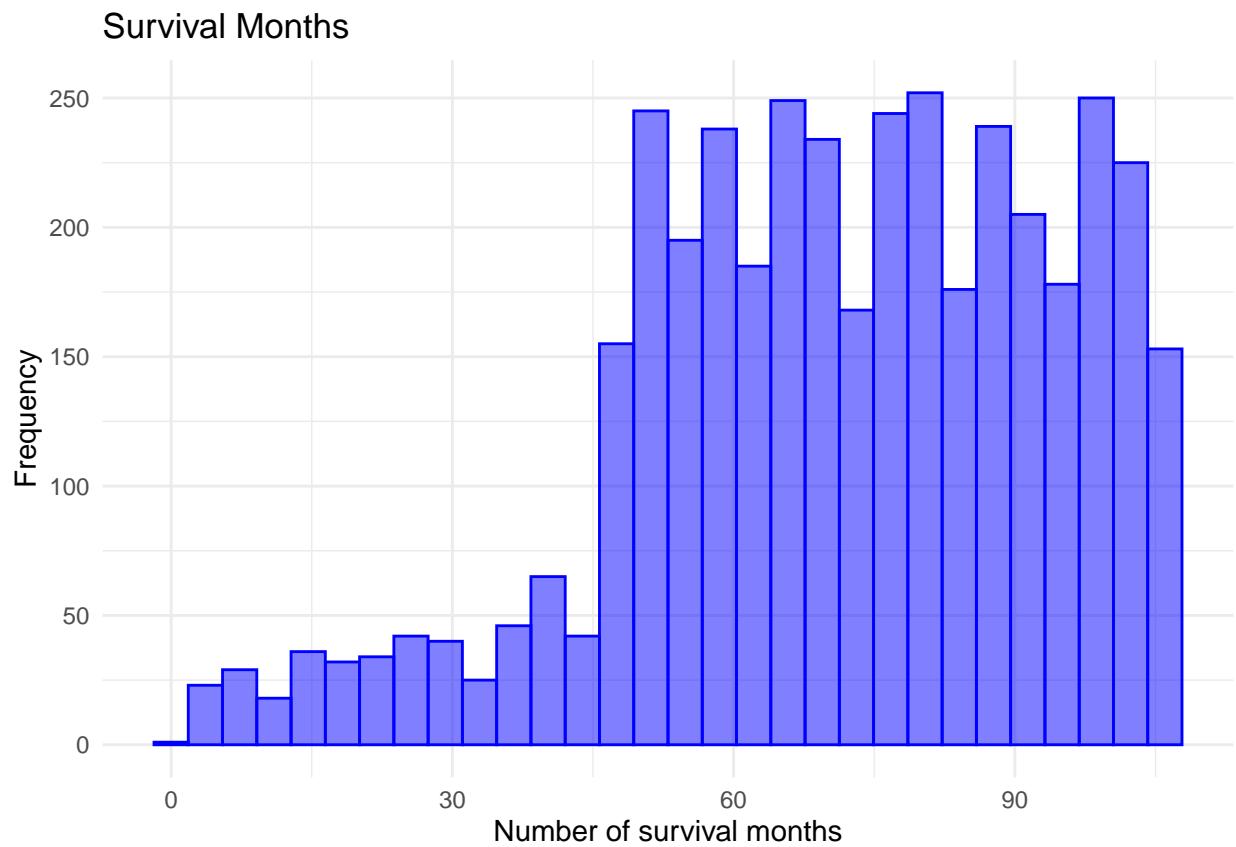
Y1 = survive months; (numeric)

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"
)
```

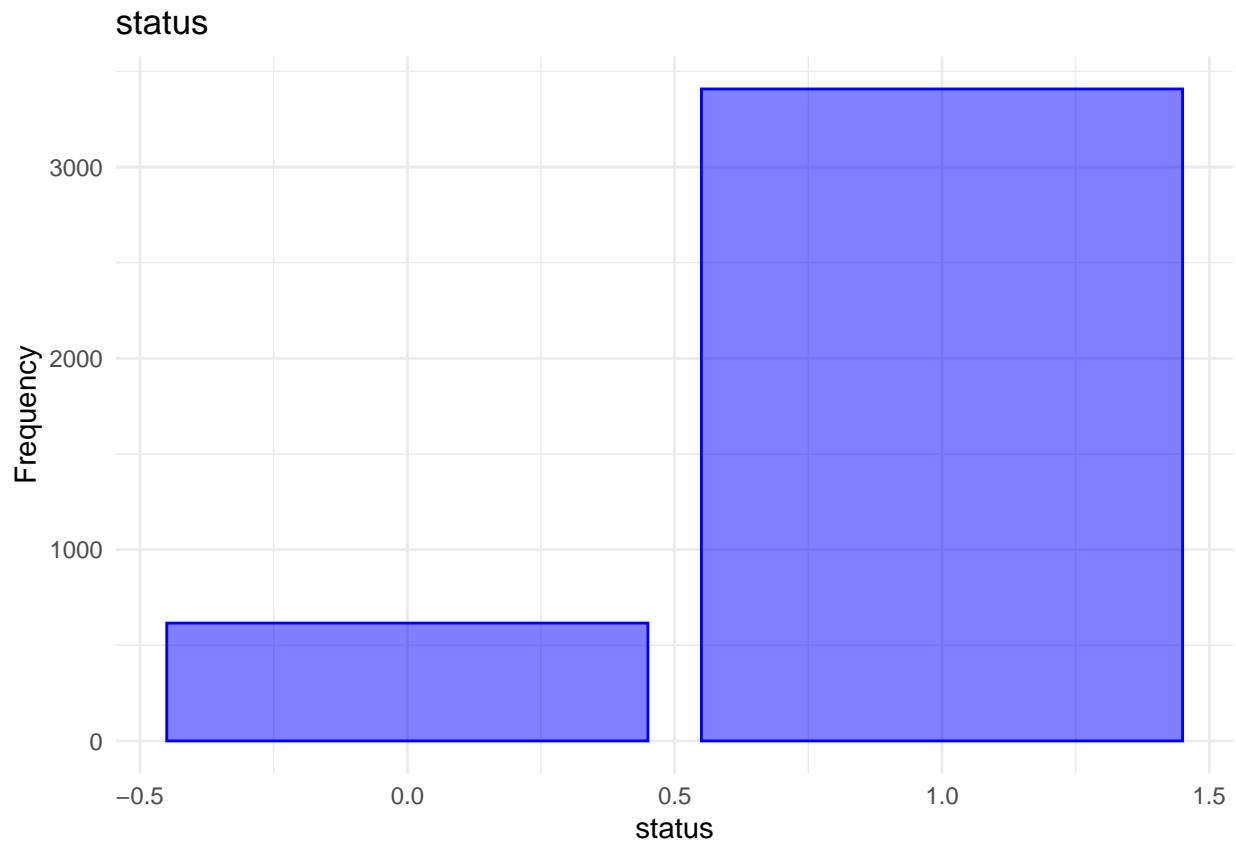
plot15survivalmonths

‘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(
```

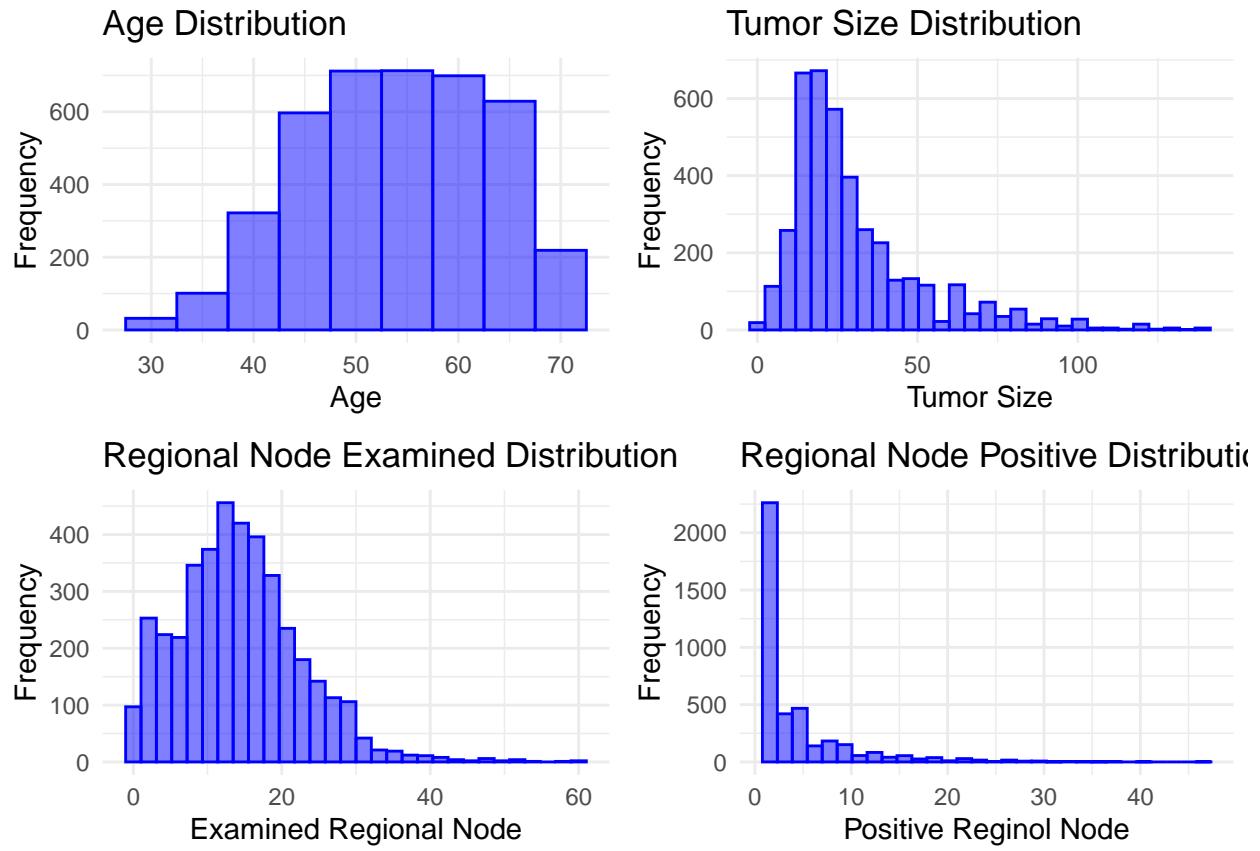
```
    title = "status",
    x = "status",
    y = "Frequency"
)
plot16status
```



Summarized plots for covariates

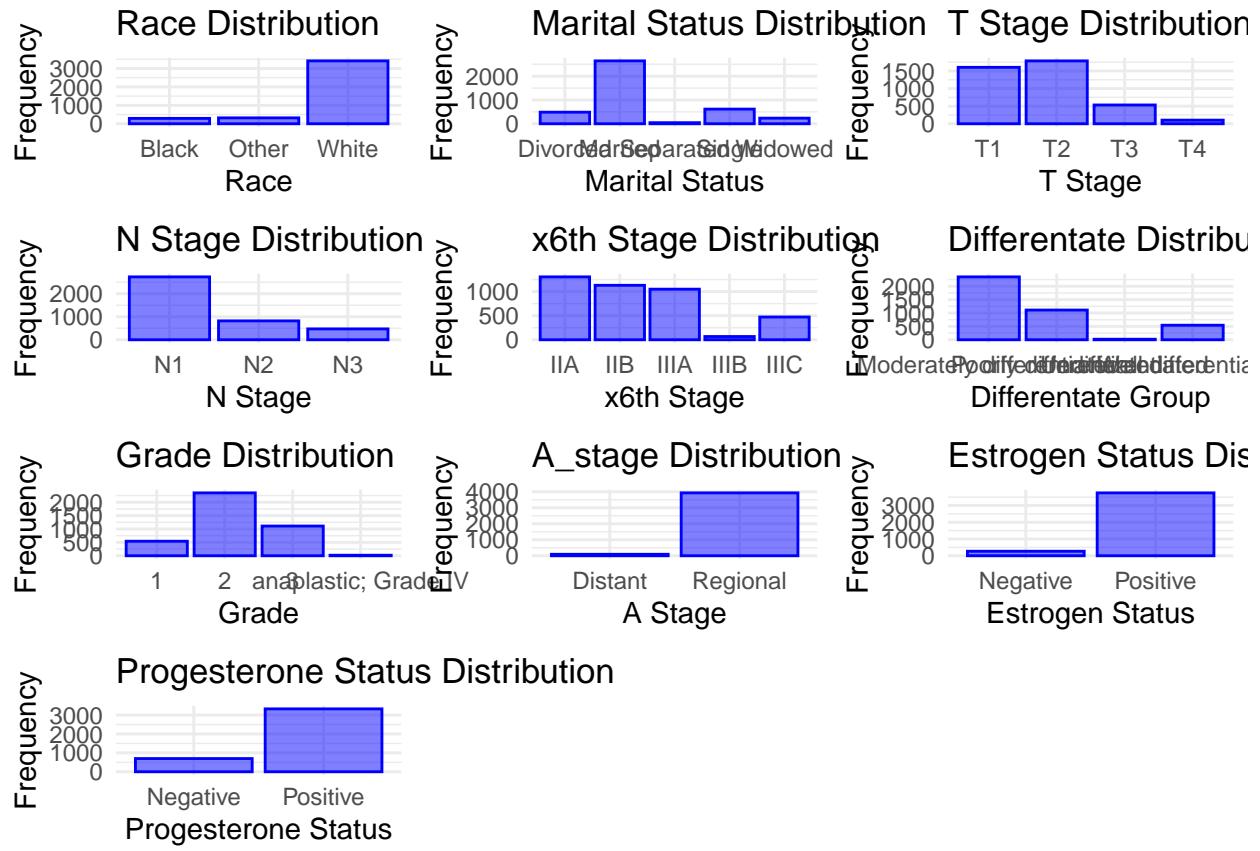
```
grid.arrange(plot1age, plot10tumorsize, plot13nodeexamined,
            plot14nodepositive, ncol = 2)
```

```
## `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, plot7differentiate,
            plot8grade, plot9astage, plot11estrogen, plot12progesterone, ncol = 3)
```

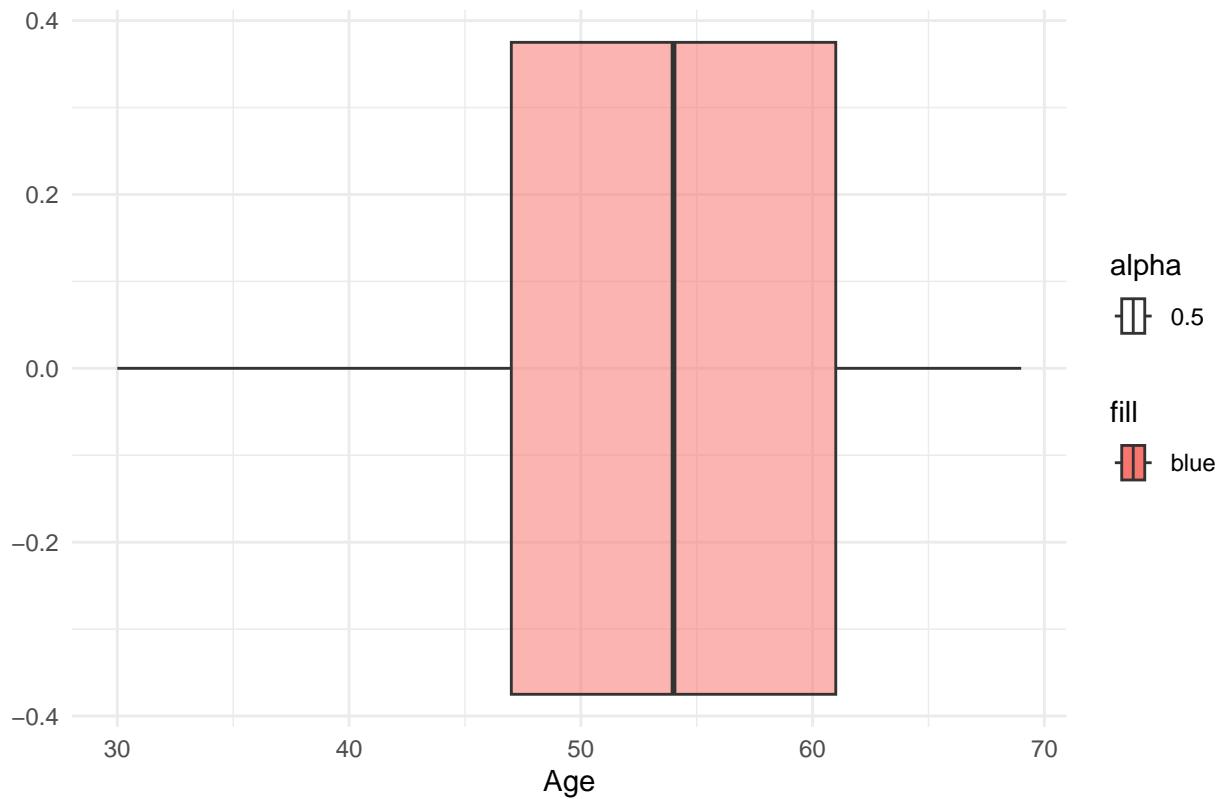


Boxplot visualisation to look for outliers

```
box_age = ggplot(bc, aes(x = age, fill = "blue", alpha = 0.5))+
  geom_boxplot() +
  theme(legend.position = 'none') +
  xlab("Age") +
  ggtitle("Box Plot Distribution of Age") +
  theme_minimal()

box_age
```

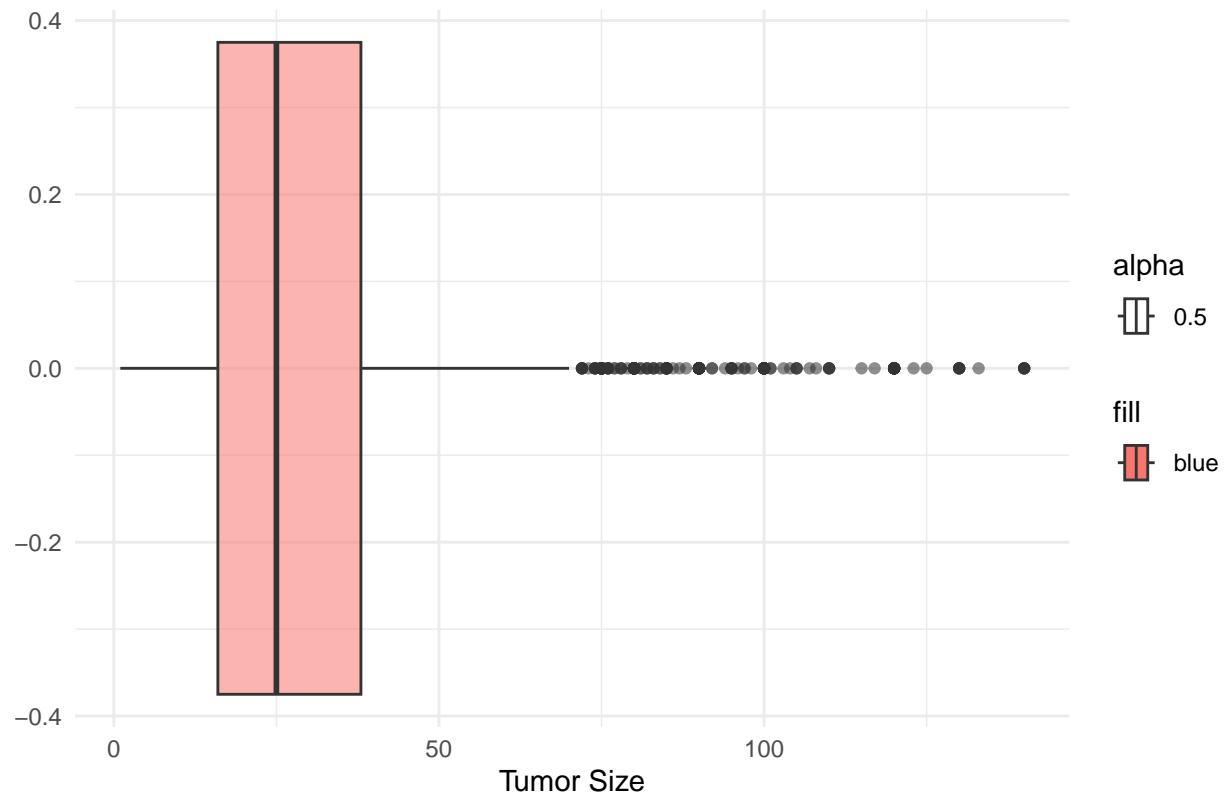
Box Plot Distribution of Age



```
box_tumor = ggplot(bc, aes(x = tumor_size, fill = "blue", alpha = 0.5)) +  
  geom_boxplot() +  
  theme(legend.position = 'none') +  
  xlab("Tumor Size") +  
  ggtitle("Box Plot Distribution of Tumor Size") +  
  theme_minimal()
```

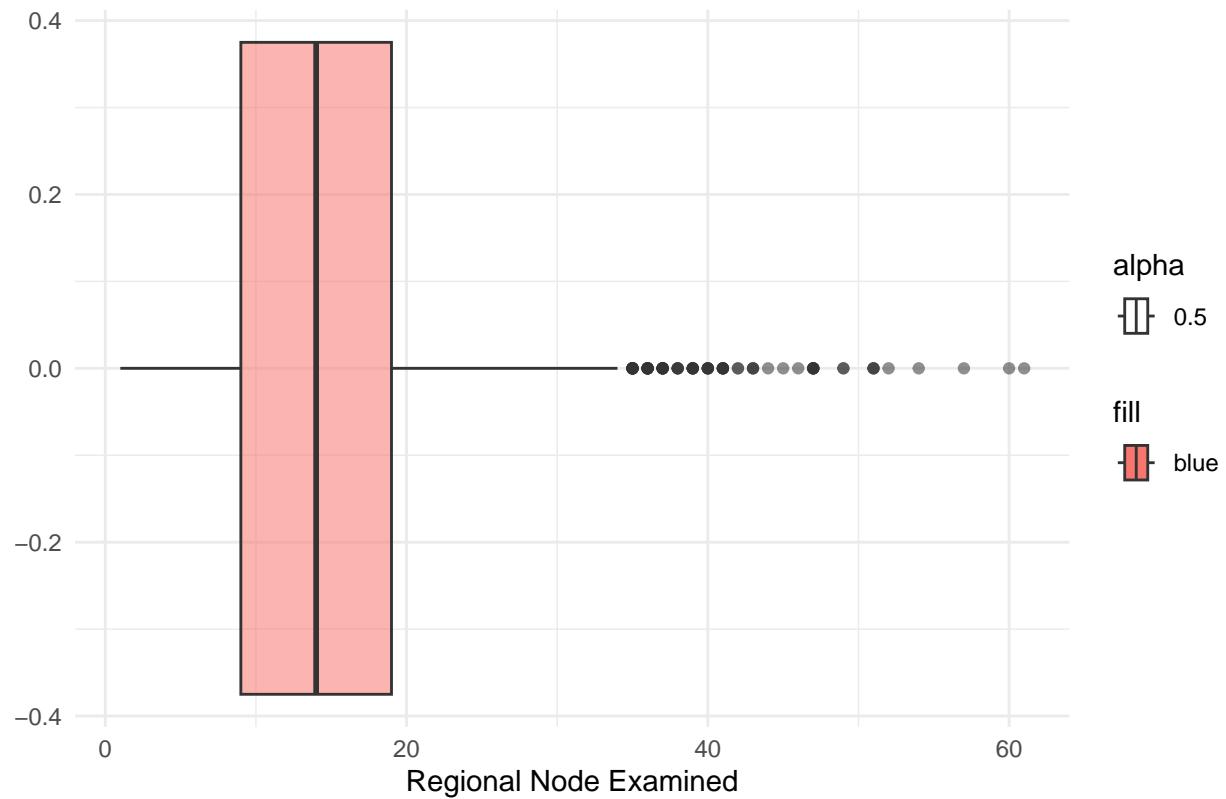
```
box_tumor
```

Box Plot Distribution of Tumor Size



```
box_examined = ggplot(bc, aes(x = regional_node_examined, fill = "blue", alpha = 0.5)) +  
  geom_boxplot() +  
  theme(legend.position = 'none') +  
  xlab("Regional Node Examined") +  
  ggtitle("Box Plot Distribution of Regional Node Examined") +  
  theme_minimal()  
  
box_examined
```

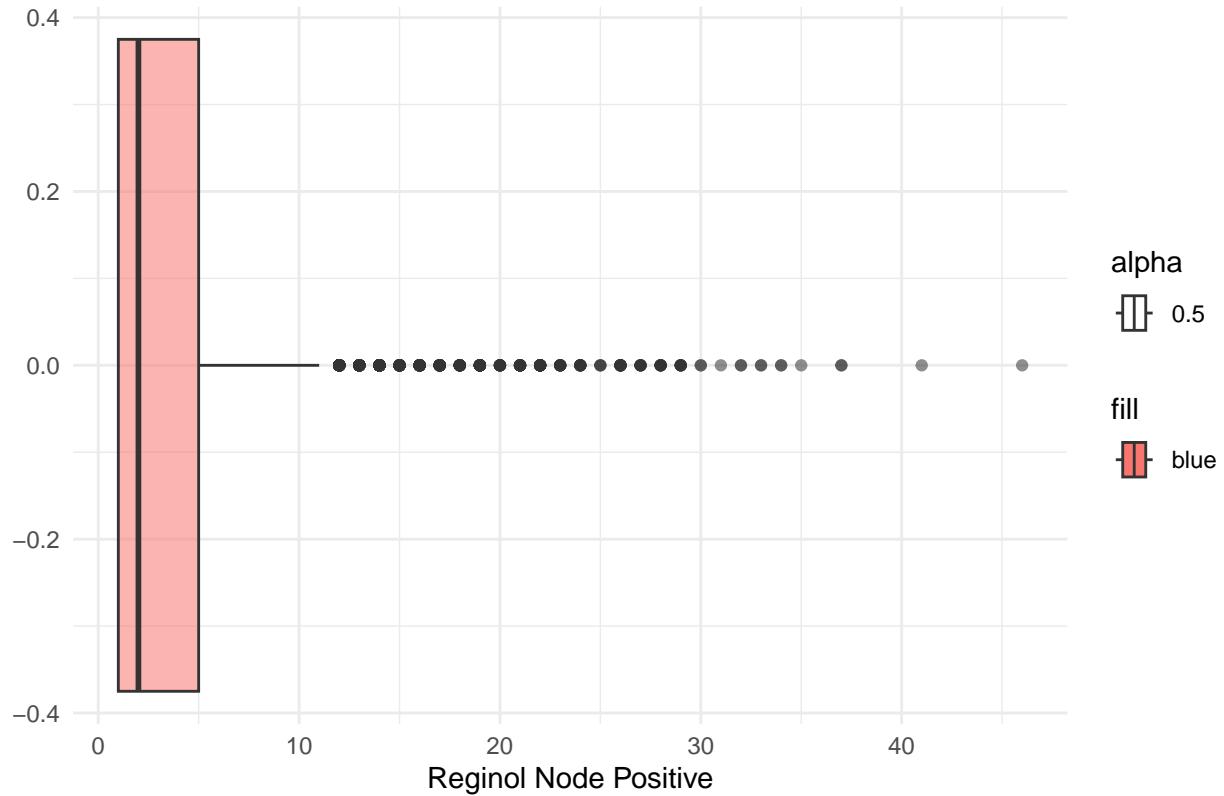
Box Plot Distribution of Regional Node Examined



```
box_positive = ggplot(bc, aes(x = reginol_node_positive, fill = "blue", alpha = 0.5))+
  geom_boxplot() +
  theme(legend.position = 'none') +
  xlab("Reginol Node Positive") +
  ggtitle("Box Plot Distribution of Reginol Node Positive") +
  theme_minimal()

box_positive
```

Box Plot Distribution of Reginol Node Positive



- From the boxplots, we discover there are outliers in all the boxplots except the age distribution, so we will remove the outliers in our next step.

Outliers Remove

```
# Find the quantile
q1 = quantile(bc$tumor_size, probs = c(0.25, 0.75), na.rm = FALSE)
iqr1 = IQR(bc$tumor_size)
upper1 = q1[2] + 1.5*iqr1
lower1 = q1[1] - 1.5*iqr1

q2 = quantile(bc$regional_node_examined, probs = c(0.25, 0.75), na.rm = FALSE)
iqr2 = IQR(bc$regional_node_examined)
upper2 = q2[2] + 1.5*iqr2
lower2 = q2[1] - 1.5*iqr2

q3 = quantile(bc$reginol_node_positive, probs = c(0.25, 0.75), na.rm = FALSE)
iqr3 = IQR(bc$reginol_node_positive)
upper3 = q3[2] + 1.5*iqr3
lower3 = q3[1] - 1.5*iqr3

bc = bc |>
  filter(
    tumor_size <= upper1 & tumor_size >= lower1
  ) |>
```

```

filter(
  regional_node_examined <= upper2 & regional_node_examined >= lower2
) |>
filter(
  reginol_node_positive <= upper3 & reginol_node_positive >= lower3
)
bc

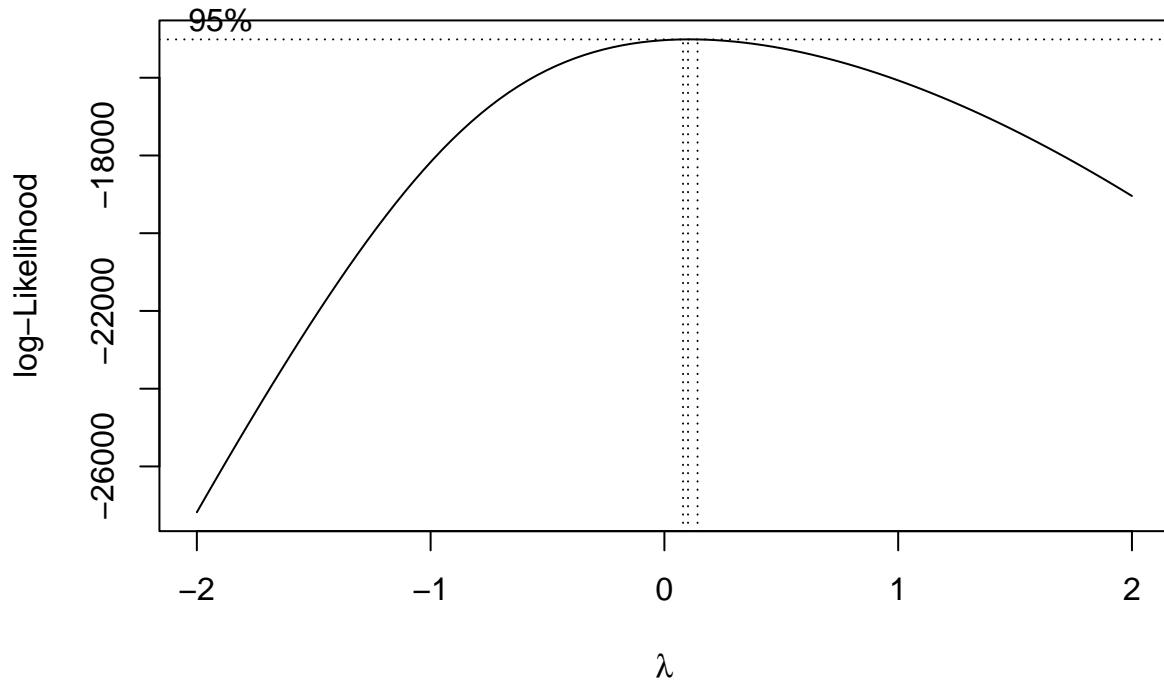
## # A tibble: 3,461 x 16
##   age race marital_status t_stage n_stage x6th_stage differentiate grade
##   <dbl> <dbl>        <dbl>    <dbl>    <dbl>        <dbl>        <dbl>    <dbl>
## 1   68     1            1        1        1            1            1        3
## 2   50     1            1        2        2            2            2        2
## 3   58     1            2        3        3            3            2        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   68     1            4        1        1            1            2        2
## 10  46     1            1        3        1            2            1        3
## # i 3,451 more rows
## # i 8 more variables: a_stage <dbl>, tumor_size <dbl>, estrogen_status <dbl>,
## #   progesterone_status <dbl>, regional_node_examined <dbl>,
## #   reginol_node_positive <dbl>, survival_months <dbl>, status <dbl>

```

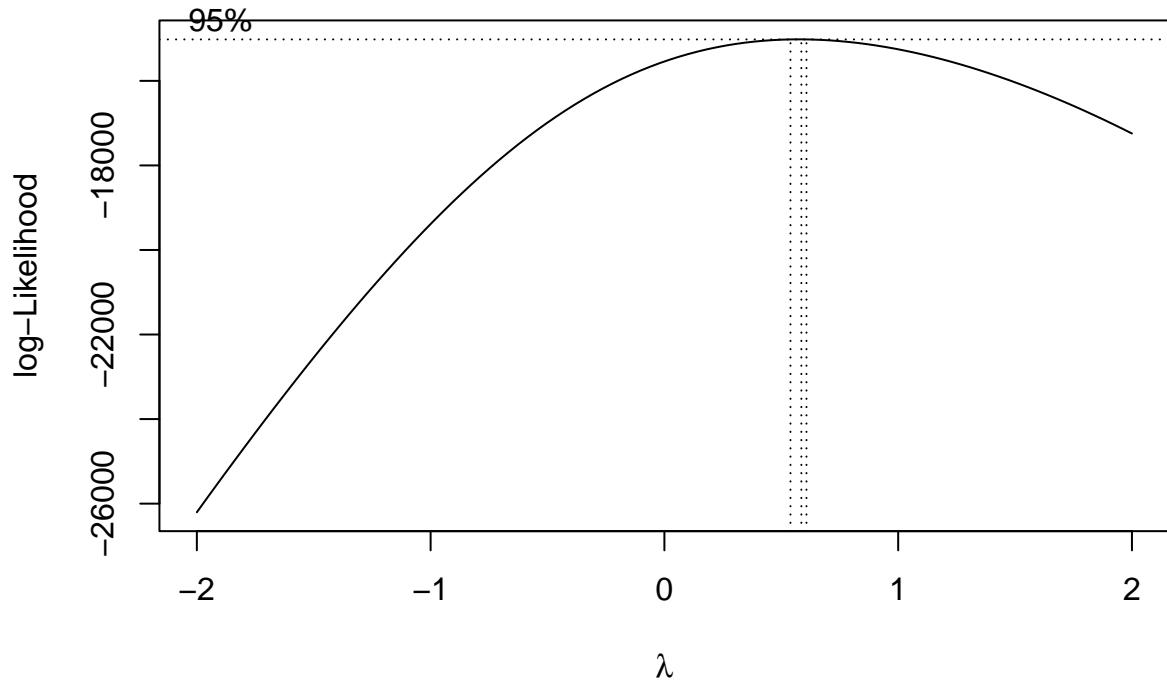
Transformation

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.

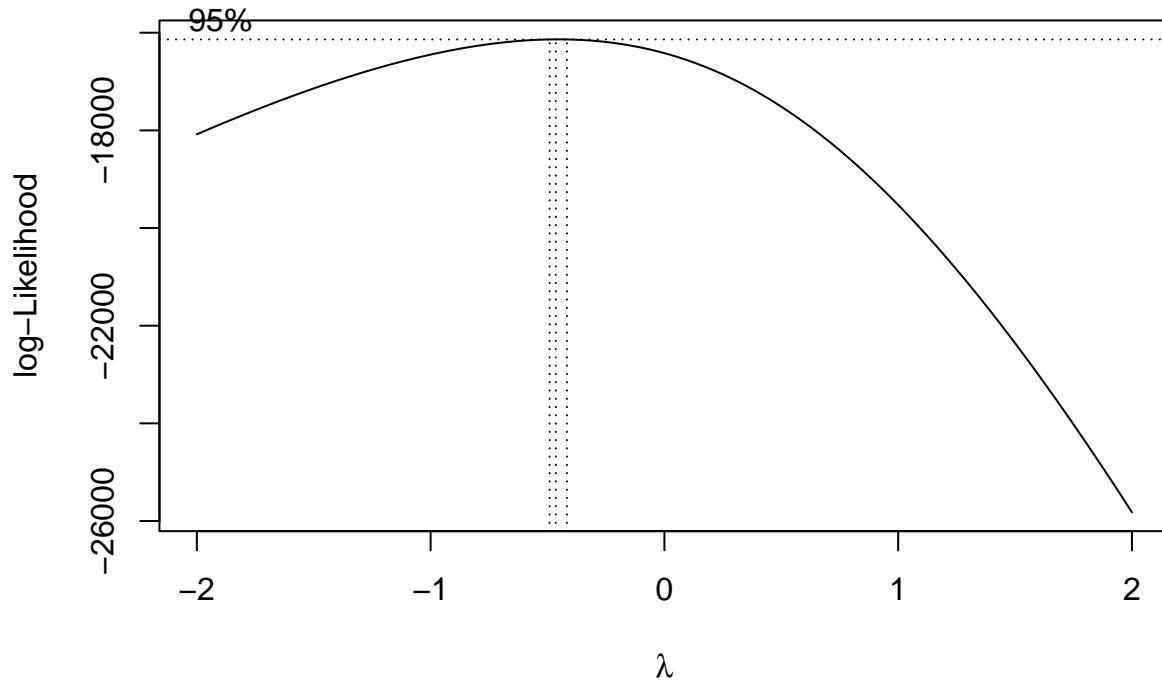
```
bc_transform_tumorsize <- boxcox(breastcancer_data$tumor_size ~ 1, lambda = seq(-2, 2, by=0.1))
```



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



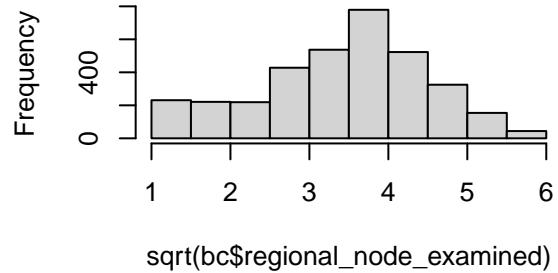
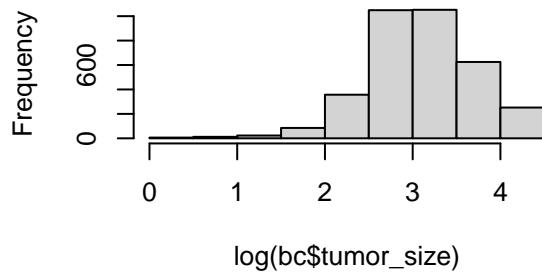
```
bc_transform_regionalnode_pos <- boxcox(breastcancer_data$reginol_node_positive ~ 1, lambda = seq(-2, 2,
```



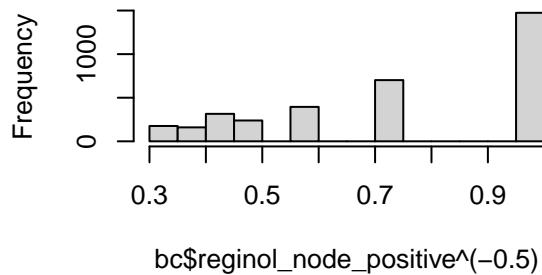
The lambda value of tumor size is close to 0, so we should use log transformation, while the lambda value of regional node examined is around 0.5, we should take a square root to the value, and the lambda value of regional node positive is around -0.5, so we should take a take square root and take an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

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

Histogram of log(bc\$tumor_size) | histogram of sqrt(bc\$regional_node_examined)



histogram of bc\$regional_node_positive^(−0.5)



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 regional_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: 3,461 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        2        2          2            2     2
## 3     58     1          2        3        3          3            2     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    68     1      4    1     1     1      2    2
## 10    46     1      1    3     1     2      1    3
## # i 3,451 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  65.86697 4.562572e-03
## X-squared1                 race  21.99791 1.671917e-05
## X-squared2                marital_status 15.09341 4.511343e-03
## X-squared3                 t_stage 41.69011 4.667993e-09
## X-squared4                 n_stage 86.19222 1.921313e-19
## X-squared5                x6th_stage 99.91678 1.024621e-20

```

```

## X-squared6      differentiate 49.08702 1.249981e-10
## X-squared7      grade        49.08702 1.249981e-10
## X-squared8      a_stage     20.66537 5.469634e-06
## X-squared9      estrogen_status 43.52930 4.176585e-11
## X-squared10     progesterone_status 72.53788 1.638562e-17
## X-squared11     reginol_node_positive 89.47869 6.794837e-15
## X-squared12      ln_tumor    99.46257 7.706226e-03
## X-squared13      sqrt_examined 37.24307 2.800207e-01

```

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)           1.801100   0.749312  2.404 0.016231 *
## race2                -0.481824   0.181257 -2.658 0.007855 **
## race3                 0.546052   0.241766  2.259 0.023909 *
## marital_status2      -0.288991   0.160997 -1.795 0.072654 .
## marital_status3      -0.146541   0.151637 -0.966 0.333847
## marital_status4      -0.152521   0.222968 -0.684 0.493943
## marital_status5      -0.964852   0.426814 -2.261 0.023785 *
## t_stage2              -0.156440   0.257180 -0.608 0.542996
## t_stage3              -0.206208   0.379434 -0.543 0.586811
## t_stage4              0.200107   0.883657  0.226 0.820849
## n_stage2              -0.361974   0.342189 -1.058 0.290138
## n_stage3              -0.741041   0.429719 -1.724 0.084621 .
## x6th_stage2            -0.150844   0.372970 -0.404 0.685890
## x6th_stage3                  NA          NA          NA          NA
## x6th_stage4            -0.276250   0.261285 -1.057 0.290386

```

```

## x6th_stage5      -0.915541  0.958406 -0.955 0.339439
## differentiate2   0.296285  0.120380  2.461 0.013845 *
## differentiate3   0.879937  0.216335  4.067 4.75e-05 ***
## differentiate4   -0.933081  0.640835 -1.456 0.145382
## grade2            NA        NA        NA        NA
## grade3            NA        NA        NA        NA
## grade4            NA        NA        NA        NA
## a_stage1          0.768184  0.437966  1.754 0.079434 .
## estrogen_status1  0.523089  0.209118  2.501 0.012370 *
## progesterone_status1  0.644165  0.140918  4.571 4.85e-06 ***
## ln_tumor           -0.099185  0.165585 -0.599 0.549175
## sqrt_examined     0.210107  0.054403  3.862 0.000112 ***
## reginol_node_positive -0.086809  0.037939 -2.288 0.022131 *
## age                -0.027971  0.006466 -4.326 1.52e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 2613.3 on 3460 degrees of freedom
## Residual deviance: 2372.5 on 3436 degrees of freedom
## AIC: 2422.5
##
## 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 logistic 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)             1.669820  0.742257  2.250 0.024471 *
## race2                  -0.483041  0.181268 -2.665 0.007704 **
## race3                  0.545443  0.241632  2.257 0.023988 *
## marital_status2        -0.288452  0.161039 -1.791 0.073262 .

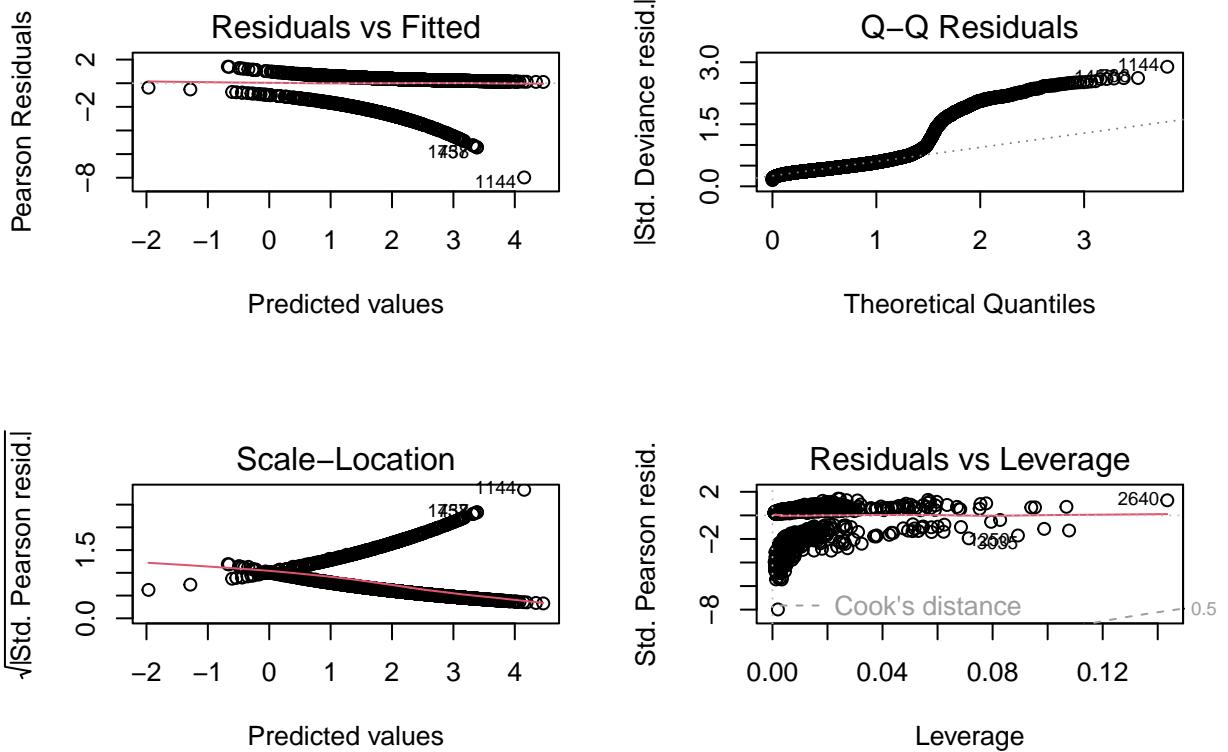
```

```

## marital_status3      -0.143745  0.151333 -0.950 0.342186
## marital_status4     -0.144722  0.223020 -0.649 0.516391
## marital_status5     -0.955754  0.428220 -2.232 0.025620 *
## t_stage2            -0.359156  0.178939 -2.007 0.044734 *
## t_stage3            -0.347678  0.306686 -1.134 0.256936
## t_stage4            -0.542171  0.429395 -1.263 0.206719
## n_stage2            -0.316154  0.195060 -1.621 0.105060
## n_stage3            -0.469552  0.371061 -1.265 0.205716
## differentiate2      0.301445  0.120242  2.507 0.012177 *
## differentiate3      0.883056  0.215825  4.092 4.29e-05 ***
## differentiate4      -0.832225  0.643368 -1.294 0.195823
## a_stage1            0.810269  0.437436  1.852 0.063981 .
## estrogen_status1    0.527199  0.209068  2.522 0.011680 *
## progesterone_status1 0.645038  0.140768  4.582 4.60e-06 ***
## ln_tumor             -0.087583  0.164874 -0.531 0.595272
## sqrt_examined       0.208411  0.054319  3.837 0.000125 ***
## reginol_node_positive -0.090653  0.037884 -2.393 0.016716 *
## age                  -0.027706  0.006456 -4.292 1.77e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 2613.3  on 3460  degrees of freedom
## Residual deviance: 2374.3  on 3439  degrees of freedom
## AIC: 2418.3
##
## 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)  1.9636    0.0562  34.936 < 2e-16 ***
## race2       -0.6265    0.1656  -3.782 0.000155 ***
## race3        0.5410    0.2338   2.314 0.020680 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
## Null deviance: 2613.3 on 3460 degrees of freedom
## Residual deviance: 2592.7 on 3458 degrees of freedom
## AIC: 2598.7
## 
## Number of Fisher Scoring iterations: 5

## 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)  1.669820  0.742257  2.250 0.024471 *
## race2       -0.483041  0.181268 -2.665 0.007704 **
## race3        0.545443  0.241632  2.257 0.023988 *
## marital_status2 -0.288452  0.161039 -1.791 0.073262 .
## marital_status3 -0.143745  0.151333 -0.950 0.342186
## marital_status4 -0.144722  0.223020 -0.649 0.516391
## marital_status5 -0.955754  0.428220 -2.232 0.025620 *
## t_stage2     -0.359156  0.178939 -2.007 0.044734 *
## t_stage3     -0.347678  0.306686 -1.134 0.256936
## t_stage4     -0.542171  0.429395 -1.263 0.206719
## n_stage2     -0.316154  0.195060 -1.621 0.105060
## n_stage3     -0.469552  0.371061 -1.265 0.205716
## differentiate2  0.301445  0.120242  2.507 0.012177 *
## differentiate3  0.883056  0.215825  4.092 4.29e-05 ***
## differentiate4 -0.832225  0.643368 -1.294 0.195823
## a_stage1      0.810269  0.437436  1.852 0.063981 .
## estrogen_status1  0.527199  0.209068  2.522 0.011680 *
## progesterone_status1 0.645038  0.140768  4.582 4.60e-06 ***
## reginol_node_positive -0.090653  0.037884 -2.393 0.016716 *
## ln_tumor       -0.087583  0.164874 -0.531 0.595272

```

```

## sqrt_examined      0.208411   0.054319   3.837 0.000125 ***
## age                 -0.027706   0.006456  -4.292 1.77e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 2613.3 on 3460 degrees of freedom
## Residual deviance: 2374.3 on 3439 degrees of freedom
## AIC: 2418.3
##
## Number of Fisher Scoring iterations: 5

## Run a logistic regression 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)                  2.522726   0.860067   2.933 0.00336 **
## race2                     -3.939061   2.157982  -1.825 0.06795 .
## race3                      1.317146  678.517674   0.002 0.99845
## marital_status2            -0.301688   0.172130  -1.753 0.07966 .
## marital_status3            0.089754   0.179356   0.500 0.61678
## marital_status4            0.116965   0.266131   0.440 0.66030
## marital_status5            -1.052417   0.507443  -2.074 0.03808 *
## t_stage2                   -0.126953   0.202998  -0.625 0.53171
## t_stage3                   -0.011254   0.349032  -0.032 0.97428
## t_stage4                   -0.506856   0.471242  -1.076 0.28212
## n_stage2                   -0.274458   0.212721  -1.290 0.19697
## n_stage3                   -0.610550   0.400474  -1.525 0.12737
## differentiate2            0.373032   0.132103   2.824 0.00475 **
## differentiate3            0.937682   0.239446   3.916 9.00e-05 ***
## differentiate4            -0.788704   0.700348  -1.126 0.26010
## a_stage1                  0.537213   0.490992   1.094 0.27389
## estrogen_status1          0.710380   0.229623   3.094 0.00198 **
## progesterone_status1       0.686130   0.153912   4.458 8.28e-06 ***
## reginol_node_positive     -0.114888   0.041361  -2.778 0.00547 **

```

```

## ln_tumor           -0.320751  0.197355 -1.625  0.10411
## sqrt_examined     0.233912  0.060035  3.896  9.77e-05 ***
## age                -0.032821  0.007161 -4.583  4.58e-06 ***
## race2:marital_status2  0.339341  0.666379  0.509  0.61059
## race3:marital_status2  -0.618404  0.742461 -0.833  0.40489
## race2:marital_status3  -1.022641  0.455715 -2.244  0.02483 *
## race3:marital_status3  -1.148739  0.757200 -1.517  0.12924
## race2:marital_status4  -0.664439  0.674543 -0.985  0.32461
## race3:marital_status4  -2.740426  0.934062 -2.934  0.00335 **
## race2:marital_status5  0.481739  1.217288  0.396  0.69229
## race3:marital_status5  -0.514287  1.442682 -0.356  0.72148
## race2:t_stage2        -0.955300  0.556300 -1.717  0.08594 .
## race3:t_stage2        -0.477768  0.883906 -0.541  0.58884
## race2:t_stage3        -0.719097  1.016740 -0.707  0.47941
## race3:t_stage3        -0.161734  1.597388 -0.101  0.91935
## race2:t_stage4        -0.235932  1.521925 -0.155  0.87680
## race3:t_stage4        12.153048 494.949345  0.025  0.98041
## race2:n_stage2        -0.420316  0.709835 -0.592  0.55376
## race3:n_stage2        -0.708345  1.163541 -0.609  0.54267
## race2:n_stage3        0.819883  1.469222  0.558  0.57682
## race3:n_stage3        11.205380 371.150096  0.030  0.97591
## race2:differentiate2 -0.360661  0.414282 -0.871  0.38399
## race3:differentiate2 -0.083304  0.561563 -0.148  0.88207
## race2:differentiate3 -0.512976  0.713404 -0.719  0.47211
## race3:differentiate3  0.241208  0.939878  0.257  0.79746
## race2:differentiate4 -1.643613  1.930490 -0.851  0.39455
## race3:differentiate4          NA          NA          NA          NA
## race2:a_stage1         2.112209  1.460836  1.446  0.14821
## race3:a_stage1         -1.623707 678.510625 -0.002  0.99809
## race2:estrogen_status1 -1.214026  0.728330 -1.667  0.09554 .
## race3:estrogen_status1  0.706925  1.505458  0.470  0.63866
## race2:progesterone_status1  0.102737  0.484289  0.212  0.83200
## race3:progesterone_status1 -2.883793  1.439992 -2.003  0.04522 *
## race2:reginol_node_positive  0.090546  0.136732  0.662  0.50783
## race3:reginol_node_positive  0.262504  0.244971  1.072  0.28391
## race2:ln_tumor          0.935650  0.410084  2.282  0.02251 *
## race3:ln_tumor          0.142697  0.888872  0.161  0.87246
## race2:sqrt_examined    -0.101194  0.183537 -0.551  0.58139
## race3:sqrt_examined    -0.230949  0.248429 -0.930  0.35256
## race2:age               0.013711  0.021734  0.631  0.52812
## race3:age               0.059960  0.029658  2.022  0.04320 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 2613.3 on 3460 degrees of freedom
## Residual deviance: 2315.3 on 3402 degrees of freedom
## AIC: 2433.3
##
## Number of Fisher Scoring iterations: 13

## Run a logistic regression 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)                 1.922110  0.782570  2.456  0.01404 *
## race2                     -1.440245  0.876932 -1.642  0.10051
## race3                      3.446428  1.895159  1.819  0.06898 .
## marital_status2            -0.298533  0.171246 -1.743  0.08128 .
## marital_status3             0.098378  0.178370  0.552  0.58127
## marital_status4             0.101948  0.264810  0.385  0.70025
## marital_status5            -1.020923  0.506070 -2.017  0.04366 *
## t_stage2                   -0.303545  0.181316 -1.674  0.09411 .
## t_stage3                   -0.212548  0.313484 -0.678  0.49776
## t_stage4                   -0.491297  0.430386 -1.142  0.25365
## n_stage2                   -0.299767  0.196352 -1.527  0.12684
## n_stage3                   -0.487390  0.372179 -1.310  0.19035
## differentiate2              0.326891  0.121192  2.697  0.00699 **
## differentiate3              0.909188  0.217906  4.172 3.01e-05 ***
## differentiate4              -0.908342  0.643389 -1.412  0.15801
## a_stage1                   0.793932  0.439372  1.807  0.07077 .
## estrogen_status1            0.558580  0.213342  2.618  0.00884 **
## progesterone_status1        0.740140  0.149420  4.953 7.29e-07 ***
## reginol_node_positive       -0.100577  0.038138 -2.637  0.00836 **
## ln_tumor                     -0.208507  0.179779 -1.160  0.24613
## sqrt_examined                0.209221  0.054680  3.826  0.00013 ***
## age                          -0.028142  0.006513 -4.321 1.55e-05 ***
## race2:marital_status2       0.430857  0.658348  0.654  0.51282
## race3:marital_status2       -0.637648  0.730974 -0.872  0.38303
## race2:marital_status3       -1.109678  0.432964 -2.563  0.01038 *
## race3:marital_status3       -1.381856  0.739709 -1.868  0.06175 .
## race2:marital_status4       -0.739699  0.620023 -1.193  0.23286
## race3:marital_status4       -1.861138  0.742128 -2.508  0.01215 *
## race2:marital_status5        0.209062  1.171629  0.178  0.85838
## race3:marital_status5       -1.325019  1.402146 -0.945  0.34466
## race2:progesterone_status1 -0.428380  0.396952 -1.079  0.28051
## race3:progesterone_status1 -2.822455  1.083674 -2.605  0.00920 **
## race2:ln_tumor                  0.534411  0.266071  2.009  0.04459 *
## race3:ln_tumor                  -0.016804  0.473176 -0.036  0.97167
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

## 
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 2613.3 on 3460 degrees of freedom
## Residual deviance: 2339.4 on 3427 degrees of freedom
## AIC: 2407.4
##
## Number of Fisher Scoring iterations: 6

```

- From the logistic regression for single race variable, we can see that the race group of black has a P-value of 6.95×10^{-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 logistic regression 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 logistic regression 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 logistic regression 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 +
##      age + t_stage + race + sqrt_examined + estrogen_status +
##      reginol_node_positive + a_stage, family = binomial, data = newbc)
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)               1.36855   0.60035  2.280 0.022631 *
## n_stage2                -0.31915   0.19476 -1.639 0.101284
## n_stage3                -0.47891   0.36906 -1.298 0.194412
## progesterone_status1    0.66437   0.14040  4.732 2.22e-06 ***
## differentiate2            0.29970   0.11986  2.500 0.012404 *
## differentiate3            0.86923   0.21430  4.056 4.99e-05 ***
## differentiate4            -0.87777   0.62808 -1.398 0.162254
## age                      -0.02735   0.00626 -4.370 1.24e-05 ***
## t_stage2                 -0.44097   0.12077 -3.651 0.000261 ***
## t_stage3                 -0.46596   0.19454 -2.395 0.016613 *
## t_stage4                 -0.64519   0.39984 -1.614 0.106611
## race2                     -0.53578   0.17667 -3.033 0.002424 **
## race3                     0.54901   0.24094  2.279 0.022691 *
## sqrt_examined             0.21147   0.05425  3.898 9.69e-05 ***
## estrogen_status1          0.49587   0.20826  2.381 0.017268 *
## reginol_node_positive     -0.09091   0.03777 -2.407 0.016085 *
## a_stage1                  0.79915   0.43703  1.829 0.067459 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 2613.3  on 3460  degrees of freedom
## Residual deviance: 2382.1  on 3444  degrees of freedom
## AIC: 2416.1
##
## Number of Fisher Scoring iterations: 5

# or summary(backward_model)
summary(backward_model)

```

```

## 
## Call:
## glm(formula = status ~ race + t_stage + differentiate + a_stage +
##      estrogen_status + progesterone_status + reginol_node_positive +
##      sqrt_examined + age, family = binomial, data = newbc)
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)               1.267444   0.589918  2.149 0.031673 *
## race2                    -0.541906   0.176208 -3.075 0.002102 **
## race3                     0.553919   0.241076  2.298 0.021579 *
## t_stage2                 -0.444318   0.120572 -3.685 0.000229 ***
## t_stage3                 -0.464382   0.194620 -2.386 0.017028 *
## t_stage4                 -0.593001   0.401139 -1.478 0.139330
## differentiate2            0.308150   0.119585  2.577 0.009972 **

```

```

## differentiate3      0.882129   0.214053   4.121 3.77e-05 ***
## differentiate4     -0.827385   0.626277  -1.321 0.186462
## a_stage1          0.922690   0.418529   2.205 0.027482 *
## estrogen_status1  0.513420   0.207524   2.474 0.013360 *
## progesterone_status1  0.662321   0.140178   4.725 2.30e-06 ***
## reginol_node_positive -0.141365   0.020450  -6.913 4.75e-12 ***
## sqrt_examined      0.211194   0.054041   3.908 9.31e-05 ***
## age                 -0.027140   0.006247  -4.345 1.40e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 2613.3  on 3460  degrees of freedom
## Residual deviance: 2384.8  on 3446  degrees of freedom
## AIC: 2414.8
##
## Number of Fisher Scoring iterations: 5

# or summary(stepwise_model)
summary(stepwise_model)

```

```

##
## Call:
## glm(formula = status ~ progesterone_status + differentiate +
##       age + t_stage + race + sqrt_examined + estrogen_status +
##       reginol_node_positive + a_stage, family = binomial, data = newbc)
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)           1.267444   0.589918   2.149 0.031673 *
## progesterone_status1  0.662321   0.140178   4.725 2.30e-06 ***
## differentiate2        0.308150   0.119585   2.577 0.009972 **
## differentiate3        0.882129   0.214053   4.121 3.77e-05 ***
## differentiate4        -0.827385   0.626277  -1.321 0.186462
## age                  -0.027140   0.006247  -4.345 1.40e-05 ***
## t_stage2             -0.444318   0.120572  -3.685 0.000229 ***
## t_stage3             -0.464382   0.194620  -2.386 0.017028 *
## t_stage4             -0.593001   0.401139  -1.478 0.139330
## race2                -0.541906   0.176208  -3.075 0.002102 **
## race3                0.553919   0.241076   2.298 0.021579 *
## sqrt_examined         0.211194   0.054041   3.908 9.31e-05 ***
## estrogen_status1     0.513420   0.207524   2.474 0.013360 *
## reginol_node_positive -0.141365   0.020450  -6.913 4.75e-12 ***
## a_stage1            0.922690   0.418529   2.205 0.027482 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 2613.3  on 3460  degrees of freedom
## Residual deviance: 2384.8  on 3446  degrees of freedom
## AIC: 2414.8
##
```

```
## 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 ~ progesterone_status + differentiate +
##       age + t_stage + race + sqrt_examined + estrogen_status +
##       reginol_node_positive + a_stage + progesterone_status:race,
##       family = binomial, data = bc_ref)
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 1.191869   0.590708  2.018  0.043623 *
## progesterone_status1        0.764857   0.148235  5.160 2.47e-07 ***
## differentiate2              0.315147   0.119838  2.630 0.008544 **
## differentiate3              0.889821   0.214405  4.150 3.32e-05 ***
## differentiate4              -0.821679   0.630015 -1.304 0.192158
## age                         -0.027579   0.006257 -4.407 1.05e-05 ***
## t_stage2                    -0.428258   0.121159 -3.535 0.000408 ***
## t_stage3                    -0.457069   0.195515 -2.338 0.019399 *
## t_stage4                    -0.575844   0.401250 -1.435 0.151252
## race2                       -0.431506   0.323287 -1.335 0.181959
## race3                       2.868574   1.023458  2.803 0.005066 **
## sqrt_examined                0.210406   0.054166  3.884 0.000103 ***
## estrogen_status1             0.556702   0.210236  2.648 0.008097 **
## reginol_node_positive        -0.144036   0.020556 -7.007 2.44e-12 ***
## a_stage1                     0.904782   0.418222  2.163 0.030510 *
## progesterone_status1:race2 -0.148944   0.384075 -0.388 0.698165
## progesterone_status1:race3 -2.779082   1.053072 -2.639 0.008315 **
## ---
```

```

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 2613.3  on 3460  degrees of freedom
## Residual deviance: 2370.7  on 3444  degrees of freedom
## AIC: 2404.7
##
## Number of Fisher Scoring iterations: 6

```

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.

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      3445    2385.3
## 2      3431    2347.0 14    38.303 0.000467 ***
## ---
## 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).

global F test for binary Y

```
anova_result1 <- anova(model_with_interaction)
print(anova_result1)

## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: status
##
## Terms added sequentially (first to last)
##
##
##          Df Deviance Resid. Df Resid. Dev
## NULL              3460    2613.3
## n_stage           2    72.918   3458    2540.4
## progesterone_status 1    55.524   3457    2484.9
## differentiate      3    25.068   3454    2459.8
## t_stage            3    16.651   3451    2443.2
## age                1    20.114   3450    2423.1
## race               2    15.795   3448    2407.3
## estrogen_status     1     4.467   3447    2402.8
## sqrt_examined       1    12.888   3446    2389.9
## reginol_node_positive 1    4.583   3445    2385.3
## marital_status      4     7.443   3441    2377.9
## progesterone_status:race 2    13.981   3439    2363.9
## race:marital_status 8    16.879   3431    2347.0

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)                 2.084674  0.442709  4.709 2.49e-06 ***
## n_stage2                  -0.341818  0.196388 -1.741 0.081767 .
## n_stage3                  -0.686330  0.360785 -1.902 0.057128 .
## progesterone_status1        0.743303  0.149216  4.981 6.31e-07 ***
## differentiate2              0.326512  0.120617  2.707 0.006789 **
## differentiate3              0.908345  0.216973  4.186 2.83e-05 ***
## differentiate4              -0.818660  0.652957 -1.254 0.209925
## t_stage2                  -0.419715  0.122195 -3.435 0.000593 ***
## t_stage3                  -0.464683  0.195605 -2.376 0.017520 *
## t_stage4                  -0.901535  0.351604 -2.564 0.010346 *
## age                      -0.027507  0.006494 -4.236 2.28e-05 ***
## race2                     0.120474  0.431259  0.279 0.779972
```

```

## race3          3.387637  1.080235  3.136 0.001713 **
## estrogen_status1    0.562233  0.213036  2.639 0.008311 **
## sqrt_examined     0.208059  0.054579  3.812 0.000138 ***
## reginol_node_positive -0.091469  0.038261 -2.391 0.016818 *
## marital_status2    -0.296665  0.170890 -1.736 0.082564 .
## marital_status3    0.086053  0.177702  0.484 0.628206
## marital_status4    0.084906  0.263468  0.322 0.747254
## marital_status5    -1.013795  0.504789 -2.008 0.044605 *
## progesterone_status1:race2 -0.351077  0.394772 -0.889 0.373833
## progesterone_status1:race3 -2.829915  1.083085 -2.613 0.008980 **
## race2:marital_status2    0.622949  0.663076  0.939 0.347483
## race3:marital_status2    -0.610023  0.729800 -0.836 0.403224
## race2:marital_status3    -1.016324  0.430383 -2.361 0.018204 *
## race3:marital_status3    -1.321156  0.737068 -1.792 0.073061 .
## race2:marital_status4    -0.703117  0.621665 -1.131 0.258046
## race3:marital_status4    -1.846153  0.739432 -2.497 0.012535 *
## race2:marital_status5    0.241389  1.201087  0.201 0.840718
## race3:marital_status5    -1.324598  1.400393 -0.946 0.344212
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 2613.3  on 3460  degrees of freedom
## Residual deviance: 2347.0  on 3431  degrees of freedom
## AIC: 2407
##
## Number of Fisher Scoring iterations: 6

```

With significant p value, we have enough evidence to show that the model performs well, 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_model1 <- stepAIC(model_with_interaction, direction = "both")
```

```

## Start:  AIC=2407.04
## 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
## - n_stage                  2   2350.9 2406.9
## <none>                      2347.0 2407.0
## - race:marital_status      8   2363.9 2407.9
## - reginol_node_positive    1   2352.7 2410.7
## - estrogen_status           1   2353.8 2411.8
## - progesterone_status:race 2   2360.5 2416.5

```

```

## - t_stage                      3  2363.2 2417.2
## - sqrt_examined                 1  2361.6 2419.6
## - age                          1  2365.4 2423.4
## - differentiate                 3  2369.9 2423.9
##
## Step: AIC=2406.95
## status ~ 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>                         2350.9 2406.9
## + n_stage                       2   2347.0 2407.0
## - race:marital_status          8   2367.6 2407.6
## - estrogen_status                1   2358.3 2412.3
## - progesterone_status:race     2   2364.8 2416.8
## - t_stage                       3   2367.4 2417.4
## - sqrt_examined                 1   2365.9 2419.9
## - age                           1   2368.9 2422.9
## - differentiate                  3   2374.2 2424.2
## - reginol_node_positive         1   2405.2 2459.2

```

```
summary(aic_selected_bin_model1)
```

```

##
## Call:
## glm(formula = status ~ 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)               2.112234  0.441696  4.782 1.73e-06 ***
## progesterone_status1      0.742211  0.148885  4.985 6.19e-07 ***
## differentiate2            0.335116  0.120301  2.786 0.005342 **
## differentiate3            0.917027  0.216672  4.232 2.31e-05 ***
## differentiate4           -0.749826  0.651454 -1.151 0.249731
## t_stage2                  -0.420982  0.121960 -3.452 0.000557 ***
## t_stage3                  -0.466637  0.195381 -2.388 0.016924 *
## t_stage4                  -0.913598  0.349981 -2.610 0.009043 **
## age                        -0.027164  0.006477 -4.194 2.74e-05 ***
## race2                      0.077486  0.430182  0.180 0.857056
## race3                      3.427965  1.082813  3.166 0.001547 **
## estrogen_status1           0.582737  0.212177  2.746 0.006024 **
## sqrt_examined              0.210595  0.054402  3.871 0.000108 ***
## reginol_node_positive      -0.155179  0.020597 -7.534 4.92e-14 ***
## marital_status2            -0.304240  0.170592 -1.783 0.074515 .
## marital_status3            0.088802  0.177395  0.501 0.616659
## marital_status4            0.080912  0.263254  0.307 0.758574
## marital_status5            -0.989818  0.504665 -1.961 0.049840 *
## progesterone_status1:race2 -0.318614  0.393475 -0.810 0.418087
## progesterone_status1:race3 -2.874920  1.084662 -2.651 0.008037 **
## race2:marital_status2      0.642112  0.660739  0.972 0.331146

```

```

## race3:marital_status2      -0.602688  0.727758 -0.828 0.407590
## race2:marital_status3     -0.999914  0.429491 -2.328 0.019905 *
## race3:marital_status3     -1.306519  0.737359 -1.772 0.076413 .
## race2:marital_status4     -0.686535  0.621456 -1.105 0.269281
## race3:marital_status4     -1.811568  0.740454 -2.447 0.014423 *
## race2:marital_status5      0.167040  1.185987  0.141 0.887993
## race3:marital_status5     -1.411976  1.399396 -1.009 0.312979
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 2613.3  on 3460  degrees of freedom
## Residual deviance: 2350.9  on 3433  degrees of freedom
## AIC: 2406.9
##
## Number of Fisher Scoring iterations: 6

bic_selected_bin_model1 <- stepAIC(model_with_interaction, direction = "both", family = binomial, k = 1)

## Start:  AIC=2591.52
## 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  2363.9 2543.2
## - n_stage                   2  2350.9 2579.1
## - t_stage                   3  2363.2 2583.3
## - progesterone_status:race  2  2360.5 2588.7
## - reginol_node_positive     1  2352.7 2589.1
## - differentiate              3  2369.9 2590.0
## - estrogen_status            1  2353.8 2590.2
## <none>                      2347.0 2591.5
## - sqrt_examined              1  2361.6 2597.9
## - age                         1  2365.4 2601.7
##
## Step:  AIC=2543.2
## 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  2371.5 2518.2
## - n_stage                     2  2367.6 2530.6
## - t_stage                     3  2380.4 2535.2
## - reginol_node_positive       1  2369.2 2540.3
## - progesterone_status:race    2  2377.9 2540.9
## - differentiate                 3  2386.1 2540.9
## - estrogen_status               1  2370.8 2542.0
## <none>                         2363.9 2543.2
## - sqrt_examined                  1  2378.2 2549.3
## - age                           1  2382.8 2554.0
## + race:marital_status           8  2347.0 2591.5

```

```

## Step: AIC=2518.22
## status ~ n_stage + progesterone_status + differentiate + t_stage +
##      age + race + estrogen_status + sqrt_examined + reginol_node_positive +
##      progesterone_status:race
##
##                                     Df Deviance    AIC
## - n_stage                         2   2375.2 2505.6
## - t_stage                          3   2388.7 2510.9
## - reginol_node_positive           1   2376.8 2515.3
## - differentiate                   3   2393.2 2515.4
## - progesterone_status:race       2   2385.3 2515.7
## - estrogen_status                 1   2377.8 2516.3
## <none>                           2371.5 2518.2
## - sqrt_examined                  1   2386.5 2525.0
## - age                            1   2391.2 2529.7
## + marital_status                 4   2363.9 2543.2
##
## Step: AIC=2505.63
## status ~ progesterone_status + differentiate + t_stage + age +
##      race + estrogen_status + sqrt_examined + reginol_node_positive +
##      progesterone_status:race
##
##                                     Df Deviance    AIC
## - t_stage                         3   2392.6 2498.6
## - differentiate                   3   2397.3 2503.2
## - progesterone_status:race       2   2389.5 2503.6
## - estrogen_status                 1   2382.0 2504.3
## <none>                           2375.2 2505.6
## - sqrt_examined                  1   2390.5 2512.7
## - age                            1   2394.6 2516.8
## + n_stage                         2   2371.5 2518.2
## + marital_status                 4   2367.6 2530.6
## - reginol_node_positive          1   2425.7 2547.9
##
## Step: AIC=2498.57
## status ~ progesterone_status + differentiate + age + race + estrogen_status +
##      sqrt_examined + reginol_node_positive + progesterone_status:race
##
##                                     Df Deviance    AIC
## - estrogen_status                 1   2399.1 2496.9
## - progesterone_status:race       2   2408.1 2497.8
## <none>                           2392.6 2498.6
## - differentiate                   3   2418.5 2500.0
## - sqrt_examined                  1   2406.7 2504.5
## + t_stage                         3   2375.2 2505.6
## - age                            1   2410.6 2508.4
## + n_stage                         2   2388.7 2510.9
## + marital_status                 4   2384.3 2522.8
## - reginol_node_positive          1   2459.1 2556.9
##
## Step: AIC=2496.88
## status ~ progesterone_status + differentiate + age + race + sqrt_examined +
##      reginol_node_positive + progesterone_status:race

```

```

##                                     Df Deviance    AIC
## - progesterone_status:race   2   2413.7 2495.2
## <none>                         2399.1 2496.9
## + estrogen_status            1   2392.6 2498.6
## - differentiate              3   2428.6 2501.9
## - sqrt_examined              1   2412.7 2502.3
## + t_stage                     3   2382.0 2504.3
## - age                          1   2415.3 2505.0
## + n_stage                     2   2394.6 2508.7
## + marital_status              4   2391.4 2521.8
## - reginol_node_positive       1   2464.3 2554.0
##
## Step:  AIC=2495.15
## status ~ progesterone_status + differentiate + age + race + sqrt_examined +
##         reginol_node_positive
##
##                                     Df Deviance    AIC
## <none>                         2413.7 2495.2
## - race                          2   2430.1 2495.3
## + progesterone_status:race     2   2399.1 2496.9
## + estrogen_status               1   2408.1 2497.8
## - differentiate                3   2442.7 2499.7
## - sqrt_examined                1   2427.5 2500.8
## + t_stage                      3   2395.4 2501.3
## - age                          1   2429.4 2502.7
## + n_stage                      2   2408.8 2506.5
## + marital_status               4   2406.0 2520.1
## - progesterone_status          1   2456.0 2529.3
## - reginol_node_positive        1   2478.0 2551.3

```

```
summary(bic_selected_bin_model1)
```

```

##
## Call:
## glm(formula = status ~ progesterone_status + differentiate +
##      age + race + sqrt_examined + reginol_node_positive, family = binomial,
##      data = bc_ref)
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)           2.131645  0.407669  5.229 1.71e-07 ***
## progesterone_status1 0.821413  0.122031  6.731 1.68e-11 ***
## differentiate2        0.383320  0.117279  3.268  0.00108 **
## differentiate3        0.970160  0.211839  4.580 4.66e-06 ***
## differentiate4       -0.889032  0.605519 -1.468  0.14205
## age                  -0.024176  0.006164 -3.922 8.77e-05 ***
## race2                 -0.537937  0.174653 -3.080  0.00207 **
## race3                 0.560248  0.241010  2.325  0.02009 *
## sqrt_examined         0.199883  0.053753  3.719  0.00020 ***
## reginol_node_positive -0.162239  0.019701 -8.235 < 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: 2613.3 on 3460 degrees of freedom
## Residual deviance: 2413.7 on 3451 degrees of freedom
## AIC: 2433.7
##
## Number of Fisher Scoring iterations: 5

```

In the interaction model, the AIC selected model is $\text{status} \sim \text{race} + \text{marital_status} + \text{t_stage} + \text{n_stage} + \text{differentiate} + \text{estrogen_status} + \text{progesterone_status} + \text{reginol_node_positive} + \ln\text{_tumor} + \sqrt{\text{_examined}} + \text{age} + \text{race:marital_status} + \text{race:progesterone_status} + \text{race:ln_tumor}$.

The BIC selected model is $\text{status} \sim \text{race} + \text{differentiate} + \text{estrogen_status} + \text{progesterone_status} + \text{reginol_node_positive} + \ln\text{_tumor} + \sqrt{\text{_examined}} + \text{age}$.

Model without Interaction Terms

```
aic_selected_bin_model2 <- stepAIC(model_no_interaction, direction = "both")
```

```

## Start: AIC=2417.34
## status ~ n_stage + progesterone_status + differentiate + t_stage +
##       age + race + estrogen_status + sqrt_examined + reginol_node_positive
##
##              Df Deviance    AIC
## <none>            2385.3 2417.3
## - n_stage          2   2389.5 2417.5
## - reginol_node_positive 1   2389.9 2419.9
## - estrogen_status  1   2390.8 2420.8
## - race             2   2401.2 2429.2
## - t_stage          3   2403.7 2429.7
## - sqrt_examined   1   2400.4 2430.4
## - differentiate    3   2406.7 2432.7
## - age              1   2404.4 2434.4
## - progesterone_status 1   2406.3 2436.3

```

```
summary(aic_selected_bin_model2)
```

```

##
## 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)           2.138220  0.428713  4.988 6.12e-07 ***
## n_stage2            -0.363094  0.194354 -1.868 0.061733 .
## n_stage3            -0.673870  0.356822 -1.889 0.058954 .
## progesterone_status1 0.663852  0.140389  4.729 2.26e-06 ***
## differentiate2       0.305799  0.119618  2.556 0.010574 *
## differentiate3       0.864405  0.214039  4.039 5.38e-05 ***

```

```

## differentiate4      -0.812407  0.638610 -1.272 0.203320
## t_stage2          -0.442975  0.120767 -3.668 0.000244 ***
## t_stage3          -0.481292  0.193867 -2.483 0.013043 *
## t_stage4          -0.958209  0.349481 -2.742 0.006110 **
## age                -0.027013  0.006254 -4.320 1.56e-05 ***
## race2              -0.534532  0.176578 -3.027 0.002469 **
## race3              0.552329  0.241117  2.291 0.021980 *
## estrogen_status1   0.491848  0.208508  2.359 0.018330 *
## sqrt_examined      0.210528  0.054200  3.884 0.000103 ***
## reginol_node_positive -0.081054  0.037808 -2.144 0.032043 *
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ',' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 2613.3  on 3460  degrees of freedom
## Residual deviance: 2385.3  on 3445  degrees of freedom
## AIC: 2417.3
##
## Number of Fisher Scoring iterations: 5

bic_selected_bin_model2 <- stepAIC(model_no_interaction, direction = "both", family = binomial, k = log

## Start:  AIC=2515.73
## status ~ n_stage + progesterone_status + differentiate + t_stage +
##         age + race + estrogen_status + sqrt_examined + reginol_node_positive
##
##                               Df Deviance    AIC
## - n_stage                  2  2389.5 2503.6
## - t_stage                  3  2403.7 2509.7
## - reginol_node_positive   1  2389.9 2512.2
## - differentiate             3  2406.7 2512.6
## - estrogen_status           1  2390.8 2513.0
## - race                      2  2401.2 2515.3
## <none>                     2385.3 2515.7
## - sqrt_examined            1  2400.4 2522.7
## - age                       1  2404.4 2526.7
## - progesterone_status       1  2406.3 2528.5
##
## Step:  AIC=2503.55
## status ~ progesterone_status + differentiate + t_stage + age +
##         race + estrogen_status + sqrt_examined + reginol_node_positive
##
##                               Df Deviance    AIC
## - t_stage                  3  2408.1 2497.8
## - differentiate             3  2411.2 2500.9
## - estrogen_status           1  2395.4 2501.3
## <none>                     2389.5 2503.6
## - race                      2  2406.0 2503.8
## - sqrt_examined            1  2404.9 2510.8
## - age                       1  2408.2 2514.2
## + n_stage                  2  2385.3 2515.7
## - progesterone_status       1  2410.5 2516.4
## - reginol_node_positive    1  2438.6 2544.6

```

```

## 
## Step: AIC=2497.75
## status ~ progesterone_status + differentiate + age + race + estrogen_status +
##      sqrt_examined + reginol_node_positive
##
##                                     Df Deviance    AIC
## - estrogen_status             1   2413.7 2495.2
## <none>                         2408.1 2497.8
## - race                          2   2424.5 2497.9
## - differentiate                3   2433.7 2498.8
## + t_stage                      3   2389.5 2503.6
## - sqrt_examined                1   2422.5 2504.0
## - age                           1   2425.3 2506.8
## + n_stage                      2   2403.7 2509.7
## - progesterone_status          1   2429.8 2511.2
## - reginol_node_positive        1   2473.7 2555.2
##
## Step: AIC=2495.15
## status ~ progesterone_status + differentiate + age + race + sqrt_examined +
##      reginol_node_positive
##
##                                     Df Deviance    AIC
## <none>                         2413.7 2495.2
## - race                          2   2430.1 2495.3
## + estrogen_status                1   2408.1 2497.8
## - differentiate                 3   2442.7 2499.7
## - sqrt_examined                 1   2427.5 2500.8
## + t_stage                       3   2395.4 2501.3
## - age                           1   2429.4 2502.7
## + n_stage                       2   2408.8 2506.5
## - progesterone_status          1   2456.0 2529.3
## - reginol_node_positive        1   2478.0 2551.3

```

```
summary(bic_selected_bin_model2)
```

```

## 
## Call:
## glm(formula = status ~ progesterone_status + differentiate +
##      age + race + sqrt_examined + reginol_node_positive, family = binomial,
##      data = newbc)
##
## Coefficients:
##                                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  2.131645  0.407669  5.229 1.71e-07 ***
## progesterone_status1         0.821413  0.122031  6.731 1.68e-11 ***
## differentiate2               0.383320  0.117279  3.268  0.00108 **
## differentiate3               0.970160  0.211839  4.580 4.66e-06 ***
## differentiate4              -0.889032  0.605519 -1.468  0.14205
## age                          -0.024176  0.006164 -3.922 8.77e-05 ***
## race2                        -0.537937  0.174653 -3.080  0.00207 **
## race3                        0.560248  0.241010  2.325  0.02009 *
## sqrt_examined                0.199883  0.053753  3.719  0.00020 ***
## reginol_node_positive       -0.162239  0.019701 -8.235 < 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: 2613.3  on 3460  degrees of freedom
## Residual deviance: 2413.7  on 3451  degrees of freedom
## AIC: 2433.7
##
## Number of Fisher Scoring iterations: 5

```

In the model without interaction, the AIC selected model is status ~ n_stage + progesterone_status + differentiate + t_stage + age + race + estrogen_status + sqrt_examined + reginol_node_positive, having a AIC value of 2961.5.

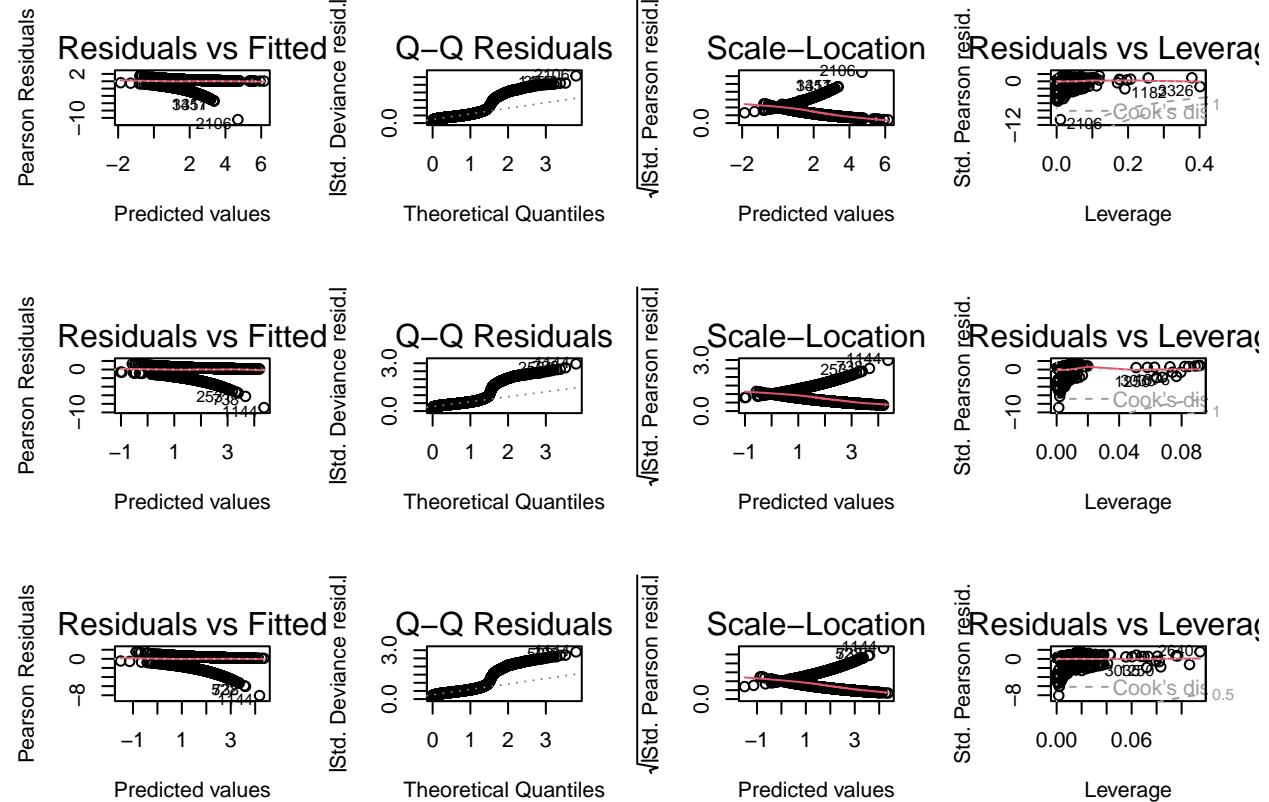
The BIC selected model is status ~ progesterone_status + differentiate + t_stage + age + race + estrogen_status + sqrt_examined + reginol_node_positive with AIC value of 3003.7.

Plots to check the model assumptions:

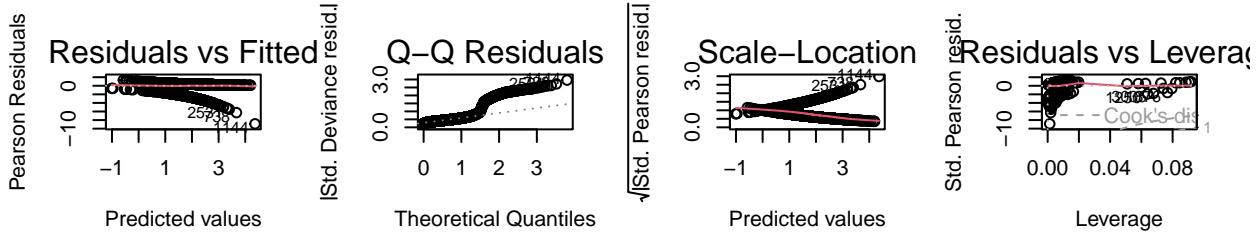
```

par(mfrow = c(3,4))
plot(aic_selected_bin_model1)
plot(bic_selected_bin_model1)
plot(aic_selected_bin_model2)

```



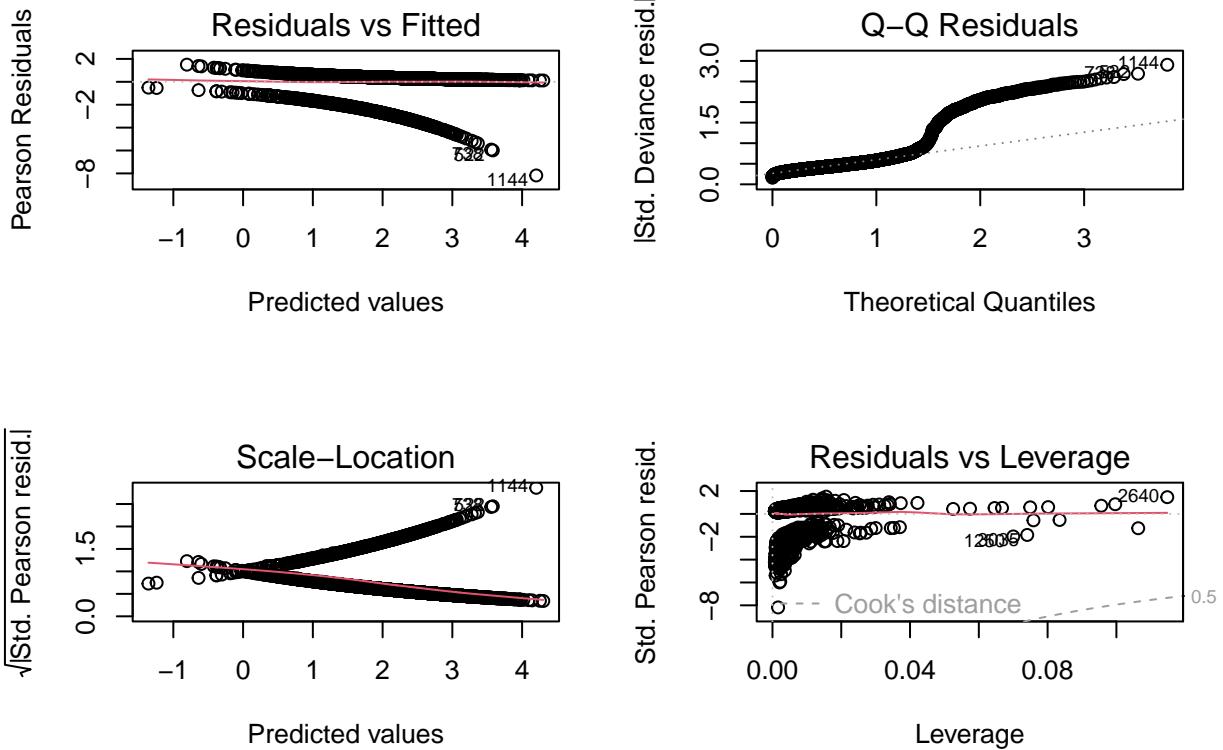
```
plot(bic_selected_bin_model2)
```



```
# final model assumption check
finalmodel <- glm(status ~ progesterone_status + differentiate + t_stage + age +
                     race + estrogen_status + sqrt_examined + reginol_node_positive,
                     data=bc_ref, family=binomial)
# check multicollinearity
vif(finalmodel)
```

```
##                                     GVIF Df GVIF^(1/(2*Df))
## progesterone_status    1.341348  1     1.158166
## differentiate          1.103290  3     1.016518
## t_stage                1.093337  3     1.014984
## age                    1.034692  1     1.017198
## race                   1.024110  2     1.005974
## estrogen_status        1.376595  1     1.173284
## sqrt_examined          1.131523  1     1.063731
## reginol_node_positive  1.178496  1     1.085586
```

```
# Model diagnostics - Plotting to check for influential outliers, equal variance
par(mfrow = c(2, 2))
plot(finalmodel)
```



```
# Compute AUC for model assessment
roc_result <- roc(bc_ref$status, fitted(finalmodel))
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
auc_result <- auc(roc_result)
print(auc_result)
```

```
## Area under the curve: 0.7119
```

For the final model: The diagnostic charts for the logistic regression model imply that the assumptions are largely met. The residuals versus fitted values plot shows a pattern that could hint at some non-linearity, but it's subtle. Few outliers are suggested by slight departures from the expected line in the tails of the normal Q-Q plot. The scale-location plot suggests consistent variance across the data. While the residuals versus leverage plot points to a few outliers, these do not appear to drastically influence the model's predictions. VIF values also suggest that the predictors in the final model are relatively independent of each other. An Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) of 0.7499 indicates that the logistic regression model has good discriminative ability.

Cross Validation

```

y <- bc_ref[["status"]] # Target column
trainIndex <- caret::createDataPartition(y, p = 0.9, list = FALSE)

# Create training and testing sets
train=bc_ref[trainIndex,]
test=bc_ref[-trainIndex,]
head(train)

## # A tibble: 6 x 15
##   age race marital_status t_stage n_stage x6th_stage differentiate grade
##   <dbl> <fct> <fct>      <fct>  <fct>  <fct>       <fct>
## 1    68 1     1           1       1       1          1       3
## 2    50 1     1           2       2       2          2       2
## 3    58 1     1           1       1       1          1       3
## 4    47 1     1           2       1       4          1       3
## 5    51 1     3           1       1       1          2       2
## 6    51 1     1           1       1       1          3       1
## # i 7 more variables: a_stage <fct>, estrogen_status <fct>,
## #   progesterone_status <fct>, reginol_node_positive <dbl>, status <dbl>,
## #   ln_tumor <dbl>, sqrt_examined <dbl>

# Fit logistic regression model with cross-validation
set.seed(123)
train_control1 <- trainControl(method = "cv", number = 10, classProbs = TRUE)
formula1 <- formula(aic_selected_bin_model1$model)
cvmodel1 <- train(formula1, data = train, method = "glm", family = "binomial",
                   trControl = train_control1)

## Warning in train.default(x, y, weights = w, ...): You are trying to do
## regression and your outcome only has two possible values Are you trying to do
## classification? If so, use a 2 level factor as your outcome column.

## Warning in train.default(x, y, weights = w, ...): cannnot compute class
## probabilities for regression

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

print(cvmodel1)

## Generalized Linear Model
##
## 3115 samples
##   9 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 2803, 2803, 2804, 2804, 2803, 2804, ...
## Resampling results:
##
##   RMSE      Rsquared      MAE
##   0.3233196  0.06097161  0.2065152

```

```

fitted_model1 <- glm(formula1, data = train, family = "binomial")
yhat1 <- predict(fitted_model1, newdata = test, type = "response")
binary_predictions1 <- ifelse(yhat1 > 0.5, 1, 0)
confusionMatrix(factor(binary_predictions1), factor(test[["status"]])))

## Confusion Matrix and Statistics
##
##             Reference
## Prediction    0     1
##           0   2   1
##           1 35 308
##
##                   Accuracy : 0.896
##                   95% CI : (0.8589, 0.9261)
##       No Information Rate : 0.8931
##       P-Value [Acc > NIR] : 0.4744
##
##                   Kappa : 0.0853
##
## McNemar's Test P-Value : 3.798e-08
##
##                   Sensitivity : 0.054054
##                   Specificity : 0.996764
##       Pos Pred Value : 0.666667
##       Neg Pred Value : 0.897959
##       Prevalence : 0.106936
##       Detection Rate : 0.005780
## Detection Prevalence : 0.008671
##       Balanced Accuracy : 0.525409
##
##       'Positive' Class : 0
##

```

RMSE = 0.3352855 Rsquared = 0.1364786 MAE = 0.2235504

Accuracy : 0.8433

```

set.seed(123)
train_control2 <- trainControl(method = "cv", number = 10, classProbs = TRUE)
formula2 <- formula(bic_selected_bin_model1$model)
cvmodel2 <- train(formula2, data = train, method = "glm", family = "binomial",
                   trControl = train_control2)

```

Warning in train.default(x, y, weights = w, ...): You are trying to do
regression and your outcome only has two possible values Are you trying to do
classification? If so, use a 2 level factor as your outcome column.

Warning in train.default(x, y, weights = w, ...): cannnot compute class
probabilities for regression

```
print(cvmodel2)
```

```

## Generalized Linear Model
##
## 3115 samples
##     6 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 2803, 2803, 2804, 2804, 2803, 2804, ...
## Resampling results:
##
##    RMSE      Rsquared      MAE
##    0.3227789  0.06145303  0.2086395

fitted_model2 <- glm(formula2, data = train, family = "binomial")
yhat2 <- predict(fitted_model2, newdata = test, type = "response")
binary_predictions2 <- ifelse(yhat1 > 0.5, 1, 0)
confusionMatrix(factor(binary_predictions2), factor(test[["status"]]))

```

```

## Confusion Matrix and Statistics
##
##             Reference
## Prediction   0     1
##           0     2     1
##           1    35   308
##
##             Accuracy : 0.896
##                 95% CI : (0.8589, 0.9261)
##     No Information Rate : 0.8931
##     P-Value [Acc > NIR] : 0.4744
##
##             Kappa : 0.0853
##
## Mcnemar's Test P-Value : 3.798e-08
##
##             Sensitivity : 0.054054
##             Specificity : 0.996764
##     Pos Pred Value : 0.6666667
##     Neg Pred Value : 0.897959
##             Prevalence : 0.106936
##             Detection Rate : 0.005780
##     Detection Prevalence : 0.008671
##             Balanced Accuracy : 0.525409
##
##     'Positive' Class : 0
##
```

RMSE = 0.3351174 Rsquared = 0.1362244 MAE = 0.2244595

Accuracy : 0.8433

```

set.seed(123)
train_control3 <- trainControl(method = "cv", number = 10, classProbs = TRUE)
formula3 <- formula(aic_selected_bin_model2$model)

```

```

cvmodel3 <- train(formula3, data = train, method = "glm", family = "binomial",
                   trControl = train_control3)

## Warning in train.default(x, y, weights = w, ...): You are trying to do
## regression and your outcome only has two possible values Are you trying to do
## classification? If so, use a 2 level factor as your outcome column.

## Warning in train.default(x, y, weights = w, ...): cannnot compute class
## probabilities for regression

print(cvmodel3)

## Generalized Linear Model
##
## 3115 samples
##     9 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 2803, 2803, 2804, 2804, 2803, 2804, ...
## Resampling results:
##
##     RMSE      Rsquared      MAE
##     0.3215198  0.06874745  0.2063091

fitted_model3 <- glm(formula3, data = train, family = "binomial")
yhat3 <- predict(aic_selected_bin_model2, newdata = test, type = "response")
binary_predictions3 <- ifelse(yhat3 > 0.5, 1, 0)
confusionMatrix(factor(binary_predictions3), factor(test[["status"]]))


## Confusion Matrix and Statistics
##
##             Reference
## Prediction   0    1
##           0    2    2
##           1   35  307
##
##                 Accuracy : 0.8931
##                 95% CI : (0.8556, 0.9236)
##     No Information Rate : 0.8931
##     P-Value [Acc > NIR] : 0.5436
##
##                 Kappa : 0.0783
##
## McNemar's Test P-Value : 1.435e-07
##
##                 Sensitivity : 0.05405
##                 Specificity : 0.99353
##     Pos Pred Value : 0.50000
##     Neg Pred Value : 0.89766
##     Prevalence : 0.10694

```

```

##          Detection Rate : 0.00578
##    Detection Prevalence : 0.01156
##    Balanced Accuracy : 0.52379
##
##    'Positive' Class : 0
##
```

RMSE = 0.3363186 Rsquared = 0.1288768 MAE = 0.2262781

Accuracy : 0.8507

```

set.seed(123)
train_control4 <- trainControl(method = "cv", number = 10, classProbs = TRUE)
formula4 <- formula(bic_selected_bin_model2$model)
cvmodel4 <- train(formula4, data = train, method = "glm", family = "binomial",
                   trControl = train_control4)
```

Warning in train.default(x, y, weights = w, ...): You are trying to do
regression and your outcome only has two possible values Are you trying to do
classification? If so, use a 2 level factor as your outcome column.

Warning in train.default(x, y, weights = w, ...): cannnot compute class
probabilities for regression

```

print(cvmodel4)
```

Generalized Linear Model

```

##
```

3115 samples

6 predictor

```

##
```

No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 2803, 2803, 2804, 2804, 2803, 2804, ...

Resampling results:

```

##
```

RMSE Rsquared MAE

0.3227789 0.06145303 0.2086395

```

fitted_model4 <- glm(formula4, data = train, family = "binomial")
yhat4 <- predict(bic_selected_bin_model2, newdata = test, type = "response")
binary_predictions4 <- ifelse(yhat4 > 0.5, 1, 0)
confusionMatrix(factor(binary_predictions4), factor(test[["status"]]))
```

Confusion Matrix and Statistics

```

##
```

Reference

| Prediction | 0 | 1 |
|------------|----|-----|
| 0 | 1 | 2 |
| 1 | 36 | 307 |

```

##
```

Accuracy : 0.8902

```

##                               95% CI : (0.8524, 0.9211)
##      No Information Rate : 0.8931
##      P-Value [Acc > NIR] : 0.6111
##
##                           Kappa : 0.0345
##
##  Mcnemar's Test P-Value : 8.636e-08
##
##                           Sensitivity : 0.027027
##                           Specificity : 0.993528
##      Pos Pred Value : 0.333333
##      Neg Pred Value : 0.895044
##                           Prevalence : 0.106936
##                           Detection Rate : 0.002890
##      Detection Prevalence : 0.008671
##                           Balanced Accuracy : 0.510277
##
##      'Positive' Class : 0
##

```

RMSE = 0.3363186 Rsquared = 0.1288768 MAE = 0.2262781

Accuracy : 0.8507

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