# Data analysis

12/09/2023

## Libraries

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

## Data Clean

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

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

```
summary(breastcancer_data)
```

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

```
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    a_stage            tumor_size     estrogen_status    progesterone_status
##  Length:4024        Min.   :  1.00   Length:4024        Length:4024
##  Class :character   1st Qu.: 16.00   Class :character   Class :character
##  Mode  :character   Median : 25.00   Mode  :character   Mode  :character
##                     Mean   : 30.47
##                     3rd Qu.: 38.00
##                     Max.   :140.00
##  regional_node_examined reginol_node_positive survival_months
##  Min.   : 1.00          Min.   : 1.000        Min.   :  1.0
##  1st Qu.: 9.00          1st Qu.: 1.000        1st Qu.: 56.0
##  Median :14.00          Median : 2.000        Median : 73.0
##  Mean   :14.36          Mean   : 4.158        Mean   : 71.3
##  3rd Qu.:19.00          3rd Qu.: 5.000        3rd Qu.: 90.0
##  Max.   :61.00          Max.   :46.000        Max.   :107.0
##     status
##  Length:4024
##  Class :character
##  Mode  :character
##
##
##
```

```
bc = breastcancer_data |>
  mutate(
    race=case_when(
      race == "White" ~ 1,
      race == "Black" ~ 2,
      race == "Other" ~ 3),
    marital_status=case_when(
      marital_status == "Married" ~ 1,
      marital_status == "Divorced" ~ 2,
      marital_status == "Single" ~ 3,
      marital_status == "Widowed" ~ 4,
      marital_status == "Separated" ~ 5),
    t_stage=case_when(
      t_stage == "T1" ~ 1,
      t_stage == "T2" ~ 2,
      t_stage == "T3" ~ 3,
      t_stage == "T4" ~ 4),
    n_stage=case_when(
      n_stage == "N1" ~ 1,
      n_stage == "N2" ~ 2,
      n_stage == "N3" ~ 3),
    x6th_stage=case_when(
      x6th_stage == "IIA" ~ 1,
      x6th_stage == "IIIA" ~ 2,
      x6th_stage == "IIIC" ~ 3,
      x6th_stage == "IIB" ~ 4,
      x6th_stage == "IIIB" ~ 5),
```

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

### Descriptive statistics for all variables

```
summary(bc)
```

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

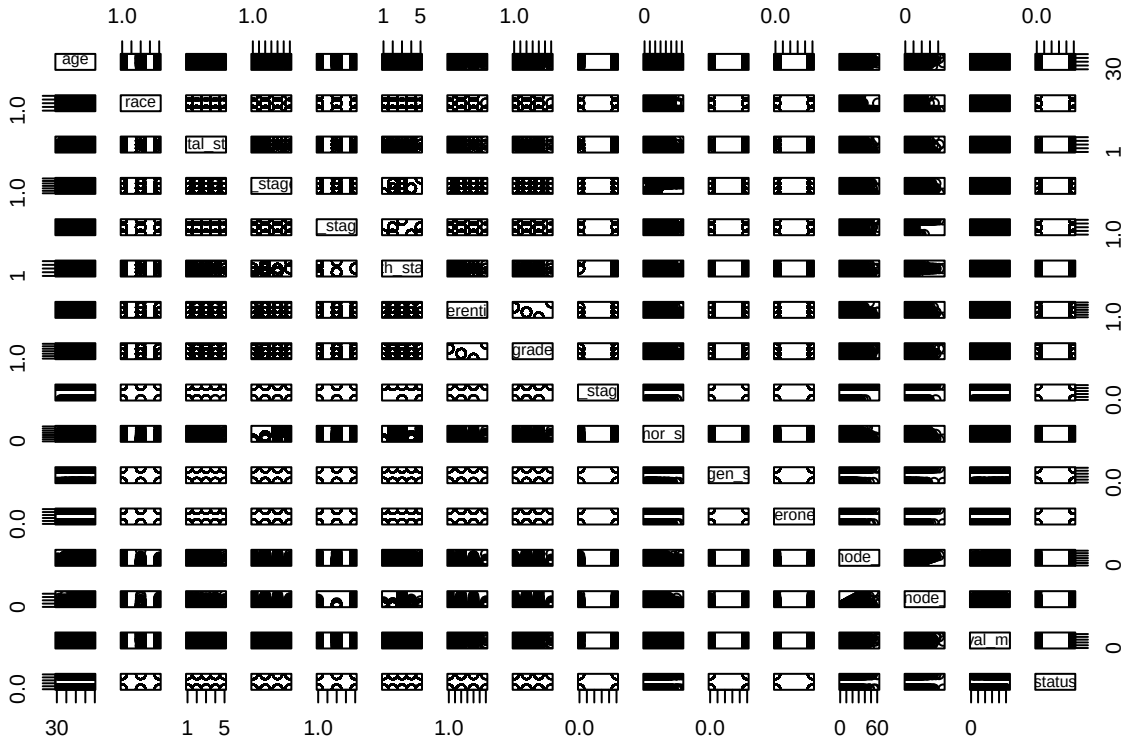
```
##  1st Qu.: 9.00          1st Qu.: 1.000        1st Qu.: 56.0   1st Qu.:1.0000
##  Median :14.00          Median : 2.000        Median : 73.0   Median :1.0000
##  Mean   :14.36          Mean   : 4.158        Mean   : 71.3   Mean   :0.8469
##  3rd Qu.:19.00          3rd Qu.: 5.000        3rd Qu.: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

We change race, marital_status, t_stage, n_stage, x6th_stage, differentiate, and grade into multiple numeric levels, while a_stage, estrogen_status, progesterone_status, and status to binary levels. The above variables are categorical variables.

And age, tumor_size, regional_node_examined, reginol_node_positive, and survival_months are numeric variables.

## Covariance and Correlation

```
plot(bc)
```




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

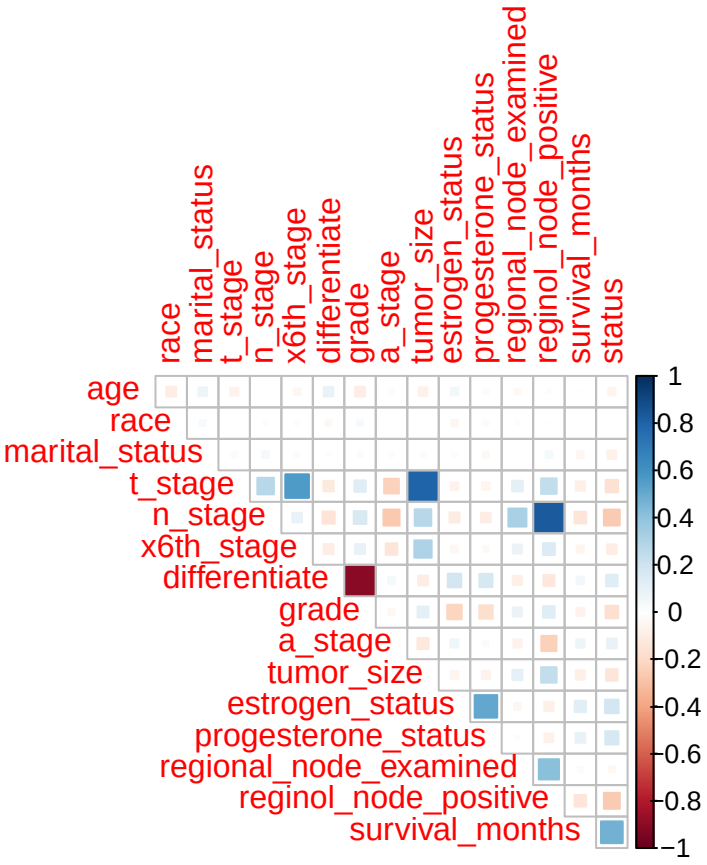
Table 1: Correlation for all variables

| | age | race | marital_status | t_stage | n_stage | x6th_stage | differentiate | grade | a_stage | tumor_size | estrogen_status | progesterone_status | regional_node_examined | reginol_node_positive | survival_months | status |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| age | 1.0000 | -0.0970 | 0.0755 | -0.0669 | 0.0029 | -0.0450 | 0.0932 | -0.0993 | 0.0209 | -0.0772 | 0.0598 | -0.0213 | -0.0126 | -0.0333 | -0.0094 | -0.0559 |

| | age | race | marital_status | t_stage | n_stage | x6th_stage | differentiate | grade | a_stage | tumor_size | estrogen_status | progesterone_status | regional_node_examined | reginol_node_positive | survival_months | status |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| race | -0.0970 | 1.0000 | 0.0349 | 0.0036 | 0.0190 | 0.0173 | -0.0336 | 0.0301 | -0.0020 | 0.0071 | -0.0425 | -0.0209 | 0.0113 | 0.0090 | -0.0025 | 0.0042 |
| marital_status | 0.0758 | 0.0349 | 1.0000 | 0.0257 | 0.0457 | 0.0221 | -0.0117 | 0.0206 | -0.0174 | 0.0207 | -0.0198 | -0.0357 | 0.0077 | 0.0443 | -0.0487 | -0.0731 |
| t_stage | -0.0669 | 0.0036 | 0.0257 | 1.0000 | 0.2770 | 0.5637 | -0.1102 | 0.1315 | -0.2211 | 0.8092 | -0.0610 | -0.0576 | 0.1141 | 0.2431 | -0.0857 | -0.1547 |
| n_stage | 0.0029 | 0.0190 | 0.0457 | 0.2770 | 1.0000 | 0.0939 | -0.1488 | 0.1625 | -0.2606 | 0.2779 | -0.1020 | -0.0937 | 0.3283 | 0.8381 | -0.1396 | -0.2558 |
| x6th_stage | -0.0450 | 0.0173 | 0.0221 | 0.5637 | 0.0939 | 1.0000 | -0.0999 | 0.0972 | -0.1372 | 0.3034 | -0.0417 | -0.0309 | 0.0826 | 0.1427 | -0.0536 | -0.0919 |
| differentiate | 0.0932 | -0.0336 | -0.0117 | -0.1102 | -0.1488 | -0.0999 | 1.0000 | -0.9083 | 0.0437 | -0.0995 | 0.1868 | 0.1758 | -0.0834 | -0.1229 | 0.0584 | 0.1342 |
| grade | -0.0993 | 0.0301 | 0.0206 | 0.1315 | 0.1625 | 0.0972 | -0.9083 | 1.0000 | -0.0395 | 0.1194 | -0.2113 | -0.1799 | 0.0844 | 0.1353 | -0.0677 | -0.1614 |
| a_stage | 0.0209 | -0.0020 | -0.0174 | -0.2211 | -0.2606 | -0.1372 | 0.0437 | -0.0395 | 1.0000 | -0.0656 | 0.1239 | 0.0265 | -0.0690 | -0.2328 | 0.0701 | 0.0966 |
| tumor_size | -0.0772 | 0.0071 | 0.0207 | 0.8092 | 0.2779 | 0.3034 | -0.0995 | 0.1194 | -0.1239 | 1.0000 | -0.0596 | -0.0699 | 0.1044 | 0.2423 | -0.0869 | -0.1342 |
| estrogen_status | 0.0598 | -0.0425 | -0.0198 | -0.0610 | -0.1020 | -0.0417 | 0.1868 | -0.2113 | 0.0656 | -0.0596 | 1.0000 | 0.5133 | -0.0448 | -0.0860 | 0.1285 | 0.1847 |
| progesterone_status | 0.0213 | -0.0209 | -0.0357 | -0.0576 | -0.0937 | -0.0309 | 0.1758 | -0.1799 | 0.0265 | -0.0699 | 0.5133 | 1.0000 | -0.0181 | -0.0781 | 0.0960 | 0.1771 |
| regional_node_examined | 0.0333 | -0.0118 | -0.0077 | -0.1141 | -0.3283 | -0.0826 | 0.0834 | -0.0844 | 0.0690 | -0.1044 | 0.0448 | 0.0181 | 1.0000 | 0.4116 | -0.0221 | -0.0348 |
| reginol_node_positive | 0.0126 | -0.0090 | -0.0443 | -0.2431 | -0.8381 | -0.1427 | 0.1229 | -0.1353 | 0.2328 | -0.2423 | 0.0860 | 0.0781 | 0.4116 | 1.0000 | -0.1352 | -0.2566 |
| survival_months | -0.0094 | 0.0025 | -0.0487 | -0.0857 | -0.1396 | -0.0536 | 0.0584 | -0.0677 | 0.0701 | -0.0869 | 0.1285 | 0.0960 | -0.0221 | -0.1352 | 1.0000 | 0.4765 |
| status | -0.0559 | 0.0042 | -0.0731 | -0.1547 | -0.2558 | -0.0919 | 0.1342 | -0.1614 | 0.0966 | -0.1342 | 0.1847 | 0.1771 | -0.0348 | -0.2566 | 0.4765 | 1.0000 |

**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
```

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

#plot7differentate
```

```r
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
```

```r
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
```

```r
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
```

```r
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
```

```r
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
```

```r
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
```

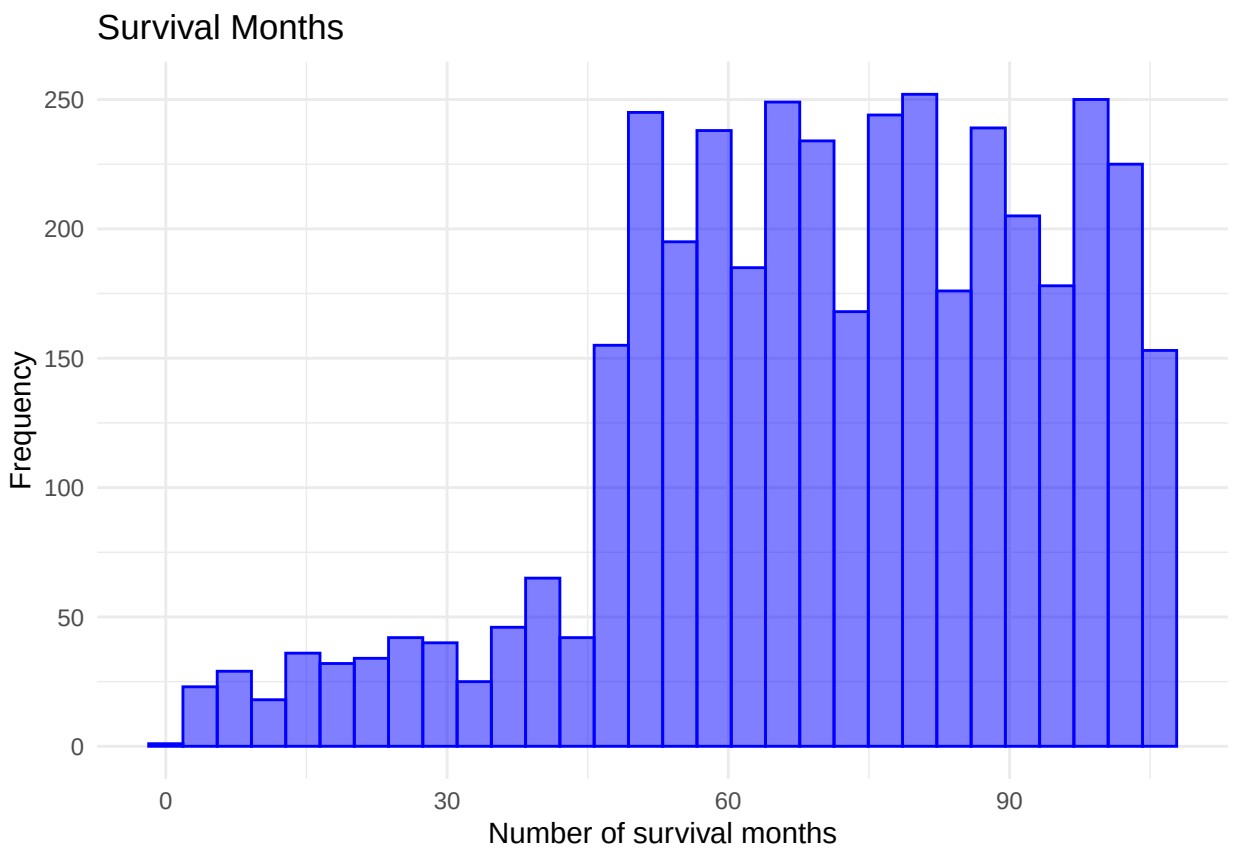
```r
plot14nodepositive =
  breastcancer_data|>
  ggplot(aes(x = reginol_node_positive)) +
  geom_histogram(color = "blue", fill = alpha("blue", 0.5))+
  theme_minimal() +
  labs(
    title = "Regional Node Positive Distribution",
    x = "Positive Reginol Node",
    y = "Frequency"
  )

#plot14nodepositive
```

Y1 = survive months; (numeric)

Y2 = status; (binary)

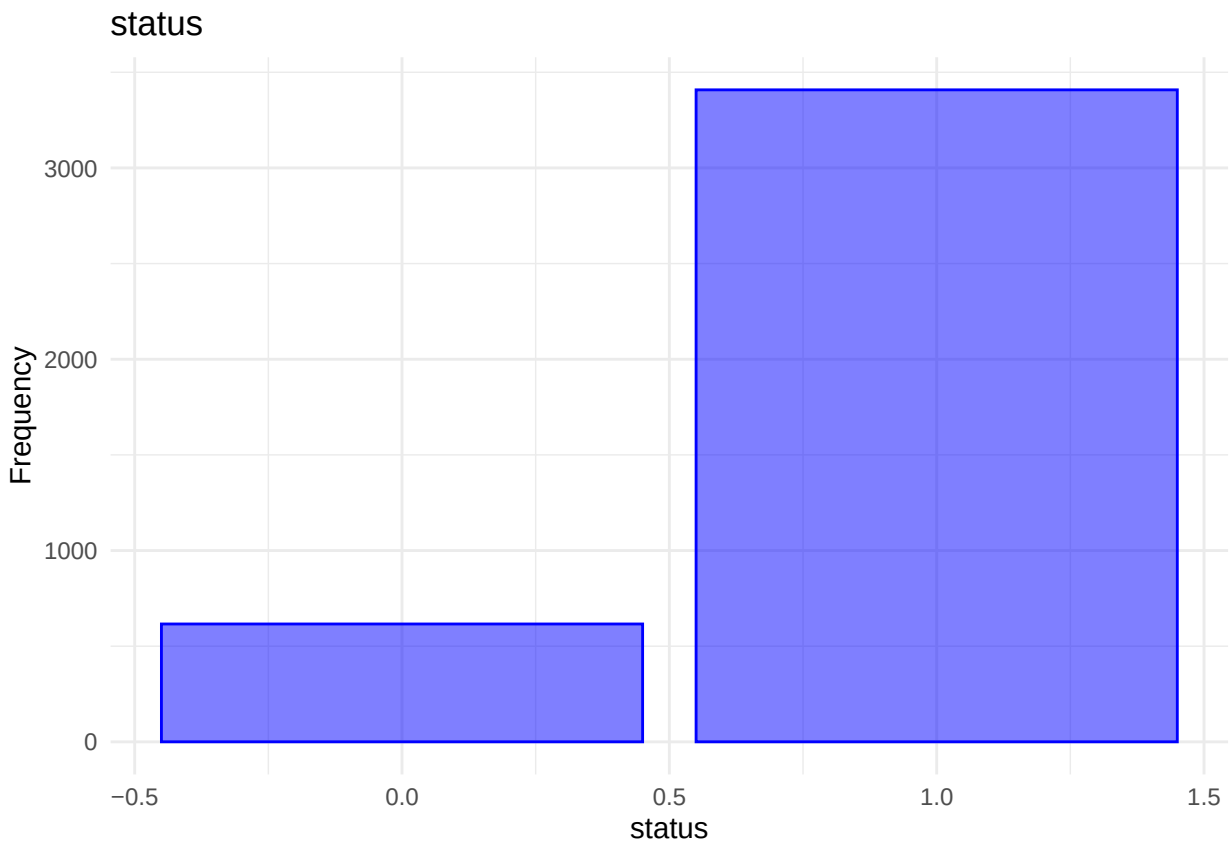
```
plot15survivalmonths =
  breastcancer_data|>
  ggplot(aes(x = survival_months)) +
  geom_histogram(color = "blue", fill = alpha("blue", 0.5))+
  theme_minimal() +
  labs(
    title = "Survival Months",
    x = "Number of survival months",
    y = "Frequency"
  )

plot15survivalmonths
```

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




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

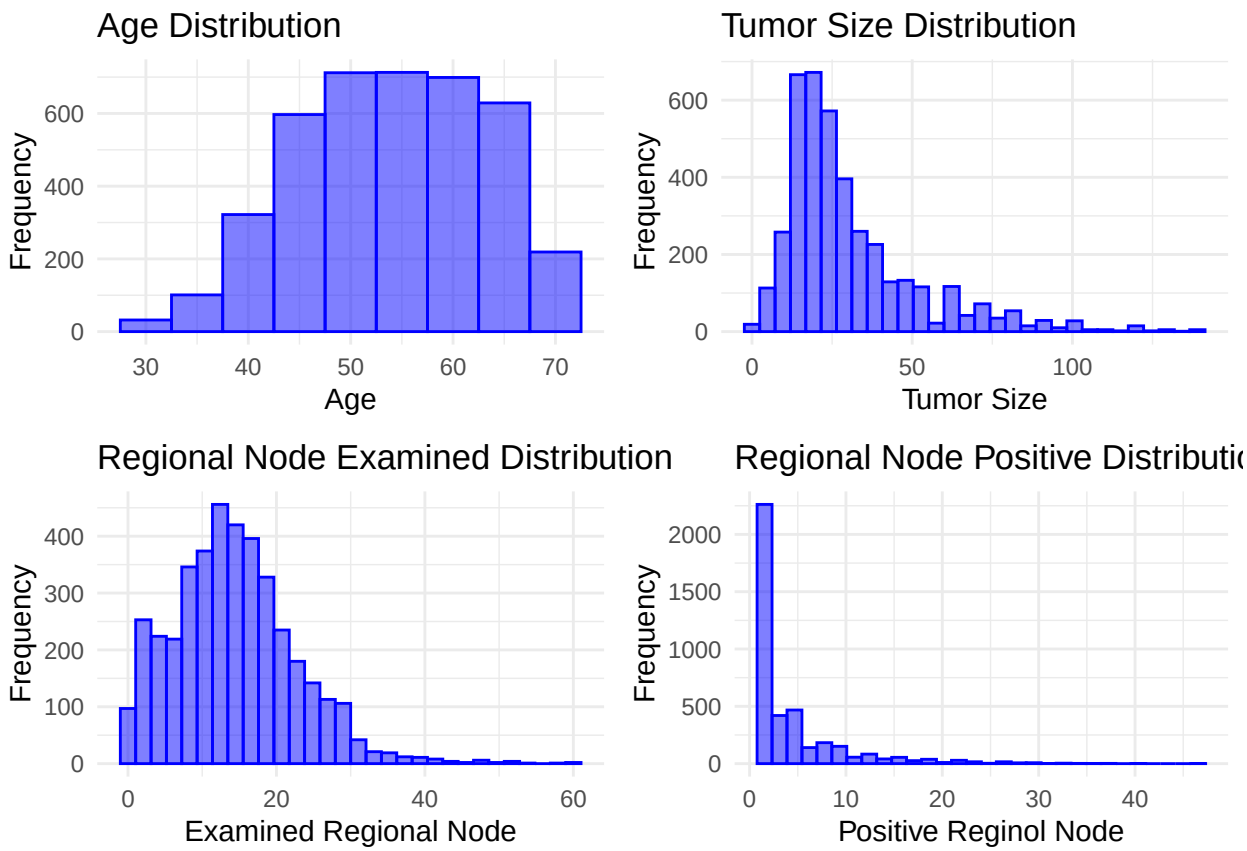
```
  )

plot16status
```




**Summarized plots for covariates**

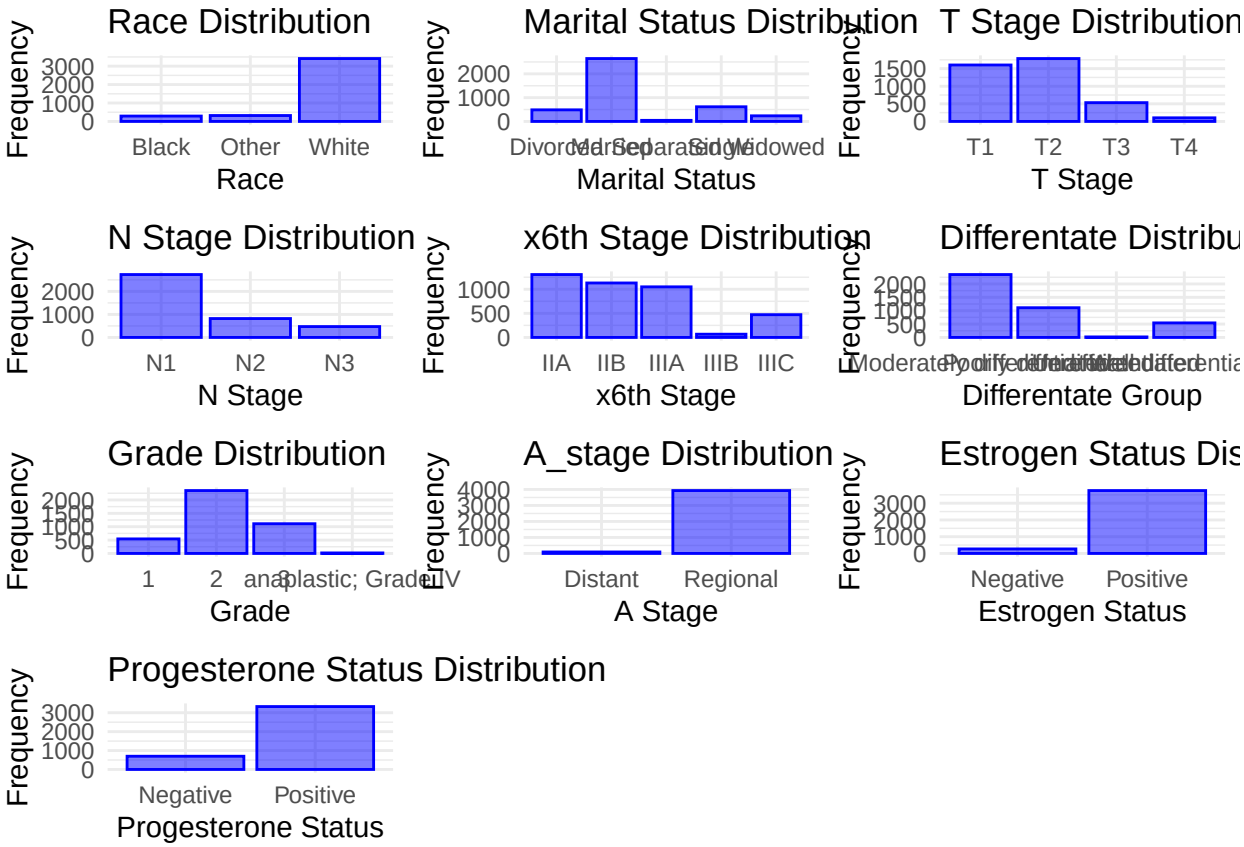
```
grid.arrange(plot1age, plot10tumorsize, plot13nodeexamined,
             plot14nodepositive, ncol = 2)
```

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




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

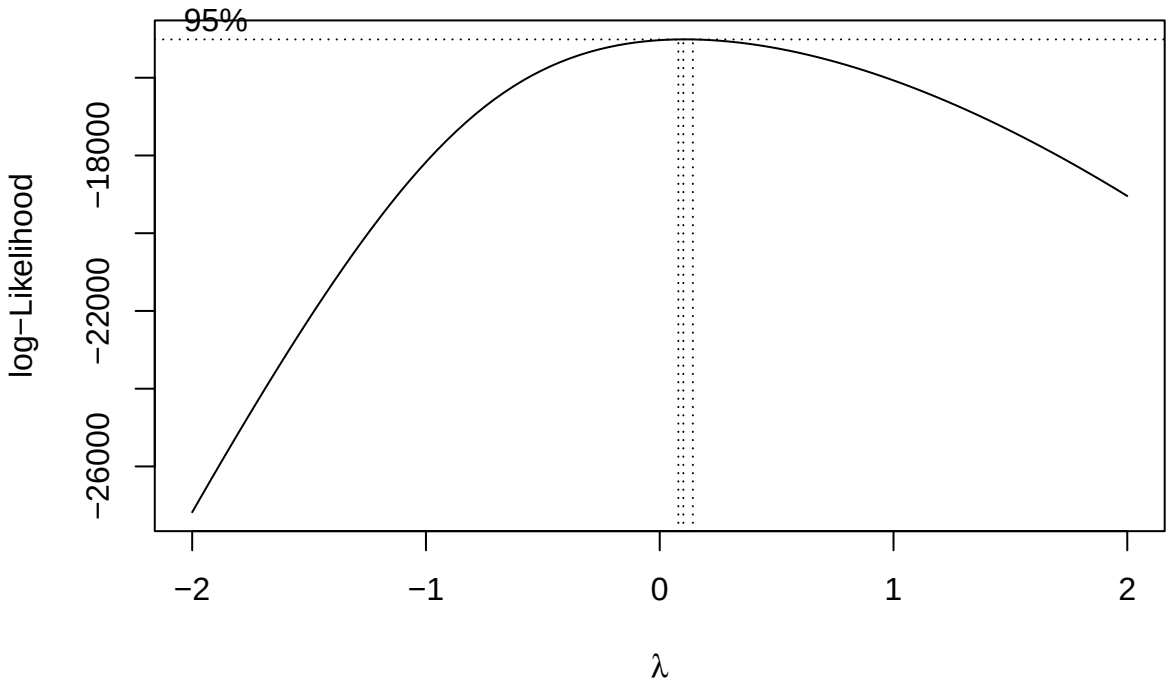
```
grid.arrange(plot2race, plot3marital,plot4tstage,
             plot5nstage, plot6x6thstage,plot7differentate,
             plot8grade,plot9astage,plot11estrogen,plot12progesterone, ncol = 3)
```



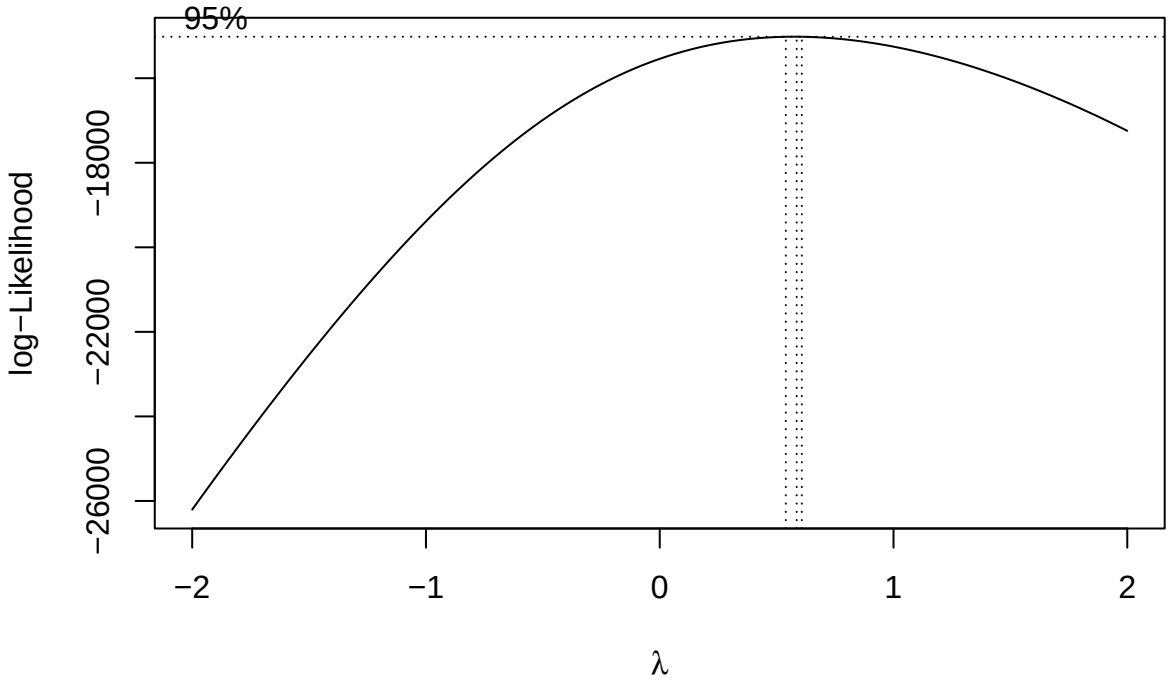

## Transformation

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

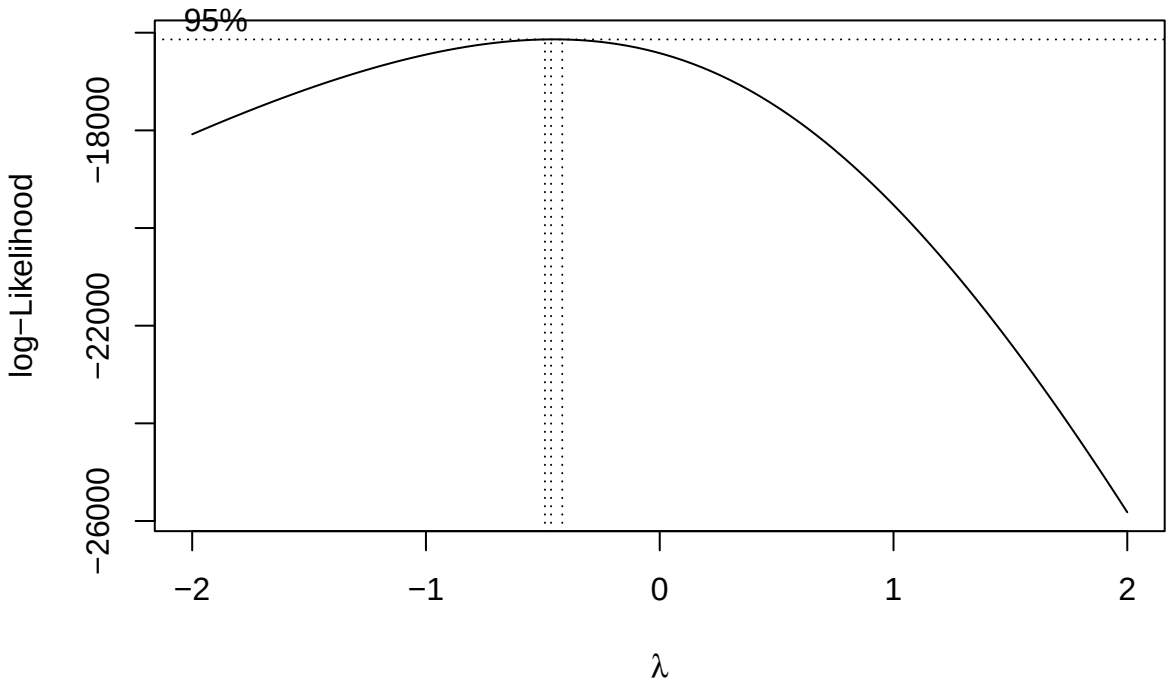
```
bc_transform_tumorsize <- boxcox(breastcancer_data$tumor_size ~ 1, lambda = seq(-2, 2, by=0.1))
```




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

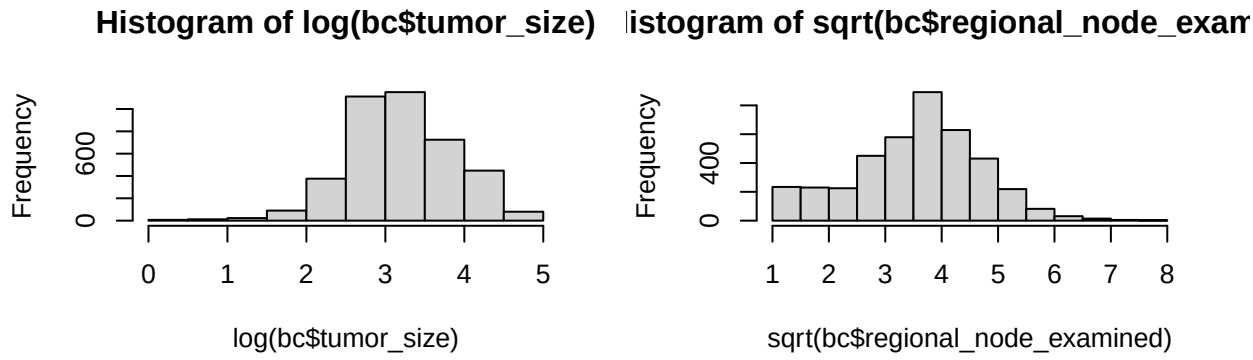


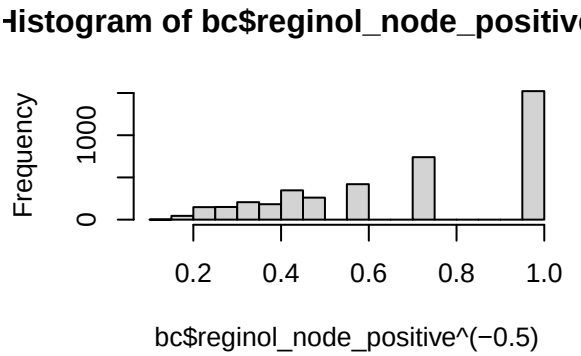

```
bc_transform_regionalnode_pos <- boxcox(breastcancer_data$reginol_node_positive ~ 1, lambda = seq(-2, 2
```




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$reginol_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 reginol_node_positive.

### Transformation model

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

```
## # A tibble: 4,024 x 15
##      age  race marital_status t_stage n_stage x6th_stage differentiate grade
##    <dbl> <dbl>          <dbl>   <dbl>   <dbl>      <dbl>          <dbl> <dbl>
##  1    68     1              1       1       1          1              1     3
##  2    50     1              1       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    40     1              2       4       3          3              1     3
## 10    69     1              1       4       3          3              3     1
```

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

## Indicator Test

### When y is status

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

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

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

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

summary(model)
```

```
##
## Call:
## glm(formula = formula, family = binomial(), data = newbc)
##
## Coefficients: (4 not defined because of singularities)
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           1.826084   0.591231   3.089  0.00201 **
## race2                -0.517829   0.161866  -3.199  0.00138 **
## race3                 0.412988   0.202323   2.041  0.04123 *
## marital_status2      -0.208737   0.141798  -1.472  0.14100
## marital_status3      -0.137065   0.134860  -1.016  0.30946
## marital_status4      -0.226852   0.192395  -1.179  0.23836
## marital_status5      -0.870139   0.369406  -2.356  0.01850 *
## t_stage2             -0.241275   0.214882  -1.123  0.26151
## t_stage3             -0.463184   0.308144  -1.503  0.13280
## t_stage4             -0.911657   0.451201  -2.021  0.04333 *
## n_stage2             -0.652606   0.238058  -2.741  0.00612 **
## n_stage3             -0.757283   0.301072  -2.515  0.01189 *
## x6th_stage2           0.069464   0.294083   0.236  0.81327
## x6th_stage3                 NA         NA      NA       NA
## x6th_stage4          -0.226947   0.231875  -0.979  0.32771
## x6th_stage5          -0.085418   0.528445  -0.162  0.87159
## differentiate2        0.387813   0.104972   3.694  0.00022 ***
## differentiate3        0.922456   0.193026   4.779 1.76e-06 ***
## differentiate4       -0.970582   0.533726  -1.819  0.06899 .
## grade2                      NA         NA      NA       NA
## grade3                      NA         NA      NA       NA
## grade4                      NA         NA      NA       NA
## a_stage1              0.044840   0.266049   0.169  0.86616
```

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

Based on the above indicator test summary, we delete grade and x6th_stage because their output was NA in the output linear model, since NA indicates these predicts may contribute collinearity.

## Model Fitting

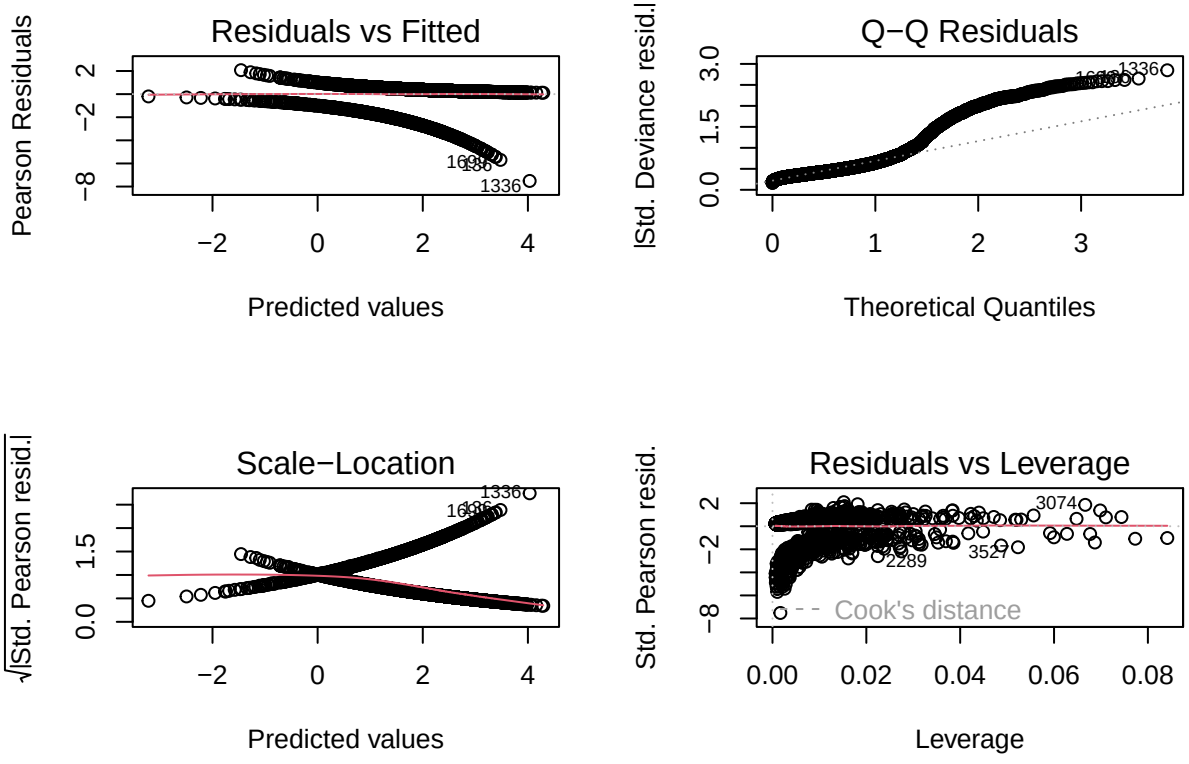
### Initial Model

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

```
##
## Call:
## glm(formula = status ~ race + marital_status + t_stage + n_stage +
##     differentiate + a_stage + estrogen_status + progesterone_status +
##     ln_tumor + sqrt_examined + reginol_node_positive + age, family = binomial,
##     data = newbc)
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)            1.775787   0.587386   3.023 0.002501 **
## race2                 -0.522765   0.161744  -3.232 0.001229 **
## race3                  0.419131   0.202312   2.072 0.038293 *
## marital_status2       -0.209325   0.141735  -1.477 0.139709
## marital_status3       -0.140584   0.134542  -1.045 0.296064
## marital_status4       -0.222113   0.192496  -1.154 0.248558
## marital_status5       -0.867369   0.370116  -2.344 0.019104 *
## t_stage2              -0.372488   0.159129  -2.341 0.019243 *
## t_stage3              -0.471905   0.266764  -1.769 0.076894 .
## t_stage4              -1.025962   0.313035  -3.277 0.001047 **
## n_stage2              -0.474091   0.129288  -3.667 0.000245 ***
## n_stage3              -0.637101   0.237639  -2.681 0.007341 **
```

```
## differentiate2          0.386027   0.104898   3.680 0.000233 ***
## differentiate3          0.917457   0.192668   4.762 1.92e-06 ***
## differentiate4         -0.947095   0.531127  -1.783 0.074557 .
## a_stage1                0.045211   0.265718   0.170 0.864894
## estrogen_status1        0.737828   0.177538   4.156 3.24e-05 ***
## progesterone_status1    0.588385   0.127499   4.615 3.93e-06 ***
## ln_tumor               -0.055274   0.138631  -0.399 0.690103
## sqrt_examined           0.255997   0.049654   5.156 2.53e-07 ***
## reginol_node_positive -0.074961   0.014984  -5.003 5.65e-07 ***
## age                   -0.023647   0.005618  -4.209 2.56e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3444.7  on 4023  degrees of freedom
## Residual deviance: 2953.4  on 4002  degrees of freedom
## AIC: 2997.4
##
## Number of Fisher Scoring iterations: 5
```

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




## Indicator Variables categorize by Race

```r
# 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
bc_ref =
  newbc |>
  mutate(
     race= relevel(
       race, ref = 1
     )
     )

bc_ref
```

```
## # A tibble: 4,024 x 15
##      age race  marital_status t_stage n_stage x6th_stage differentiate grade
##    <dbl> <fct> <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     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    40 1     2              4       3       3          1              3
## 10    69 1     1              4       3       3          3              1
## # i 4,014 more rows
## # i 7 more variables: a_stage <fct>, estrogen_status <fct>,
## #   progesterone_status <fct>, reginol_node_positive <dbl>, status <dbl>,
## #   ln_tumor <dbl>, sqrt_examined <dbl>
```

```r
## Run a SLR for race only
single_race = glm(status ~ race, family = binomial, bc_ref)

summary(single_race)
```

```
##
## Call:
## glm(formula = status ~ race, family = binomial, data = bc_ref)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.73909    0.04801  36.221  < 2e-16 ***
## race2       -0.64505    0.14350  -4.495 6.95e-06 ***
## race3        0.42389    0.18997   2.231   0.0257 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 3444.7  on 4023  degrees of freedom
## Residual deviance: 3418.9  on 4021  degrees of freedom
## AIC: 3424.9
##
## Number of Fisher Scoring iterations: 4
```

```
## Run a MLR with race ref but without interaction terms
model_ref = glm(status ~ race + marital_status + t_stage + n_stage + differentiate +
              a_stage + estrogen_status + progesterone_status + reginol_node_positive + ln_tumor +
              sqrt_examined + age,
              data = bc_ref, family = binomial)
#model_ref
summary(model_ref)
```

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

```
## Number of Fisher Scoring iterations: 5
```

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

```
##
## Call:
## glm(formula = status ~ race + marital_status + t_stage + n_stage +
##     differentiate + a_stage + estrogen_status + progesterone_status +
##     reginol_node_positive + ln_tumor + sqrt_examined + age +
##     race * (marital_status + t_stage + n_stage + differentiate +
##         a_stage + estrogen_status + progesterone_status + reginol_node_positive +
##         ln_tumor + sqrt_examined + age), family = binomial, data = bc_ref)
##
## Coefficients: (1 not defined because of singularities)
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  2.143663   0.663026   3.233 0.001224 **
## race2                       -3.268199   1.771947  -1.844 0.065123 .
## race3                        5.514563   3.122929   1.766 0.077424 .
## marital_status2             -0.183704   0.152054  -1.208 0.226988
## marital_status3              0.085036   0.158426   0.537 0.591437
## marital_status4             -0.055626   0.220307  -0.252 0.800660
## marital_status5             -0.799662   0.435399  -1.837 0.066266 .
## t_stage2                    -0.275236   0.176881  -1.556 0.119696
## t_stage3                    -0.295508   0.299330  -0.987 0.323529
## t_stage4                    -0.940666   0.342650  -2.745 0.006046 **
## n_stage2                    -0.567320   0.141733  -4.003 6.26e-05 ***
## n_stage3                    -0.843987   0.258302  -3.267 0.001085 **
## differentiate2               0.436042   0.115677   3.769 0.000164 ***
## differentiate3               1.011119   0.214291   4.718 2.38e-06 ***
## differentiate4              -0.948202   0.577004  -1.643 0.100317
## a_stage1                    -0.060677   0.293299  -0.207 0.836106
## estrogen_status1             0.806235   0.196393   4.105 4.04e-05 ***
## progesterone_status1         0.660775   0.138367   4.776 1.79e-06 ***
## reginol_node_positive       -0.065841   0.016154  -4.076 4.59e-05 ***
## ln_tumor                    -0.165689   0.159061  -1.042 0.297566
## sqrt_examined                0.278198   0.054658   5.090 3.58e-07 ***
## age                         -0.027956   0.006220  -4.495 6.97e-06 ***
## race2:marital_status2       -0.231457   0.537505  -0.431 0.666750
## race3:marital_status2       -0.493298   0.712072  -0.693 0.488458
## race2:marital_status3       -1.145440   0.409405  -2.798 0.005145 **
## race3:marital_status3       -1.209396   0.612332  -1.975 0.048261 *
## race2:marital_status4       -0.440236   0.601871  -0.731 0.464507
## race3:marital_status4       -2.132622   0.838528  -2.543 0.010981 *
## race2:marital_status5        0.098692   1.175316   0.084 0.933080
## race3:marital_status5       -0.880739   1.347639  -0.654 0.513407
```

```
## race2:t_stage2              -0.465403   0.502263  -0.927 0.354128
## race3:t_stage2               0.437497   0.782923   0.559 0.576298
## race2:t_stage3              -0.703756   0.806589  -0.873 0.382931
## race3:t_stage3               1.603080   1.333557   1.202 0.229322
## race2:t_stage4               0.014416   1.193986   0.012 0.990367
## race3:t_stage4              -0.127448   1.726137  -0.074 0.941142
## race2:n_stage2               0.484228   0.476679   1.016 0.309707
## race3:n_stage2               1.082513   0.645862   1.676 0.093724 .
## race2:n_stage3               1.463405   0.897850   1.630 0.103123
## race3:n_stage3               2.525936   1.462710   1.727 0.084188 .
## race2:differentiate2        -0.208584   0.358607  -0.582 0.560802
## race3:differentiate2        -0.109563   0.465485  -0.235 0.813919
## race2:differentiate3        -0.381274   0.682640  -0.559 0.576483
## race3:differentiate3        -0.140078   0.835274  -0.168 0.866816
## race2:differentiate4        -0.724327   1.669867  -0.434 0.664460
## race3:differentiate4               NA         NA      NA       NA
## race2:a_stage1               1.903928   1.021605   1.864 0.062369 .
## race3:a_stage1              -1.581496   1.588021  -0.996 0.319303
## race2:estrogen_status1      -0.686042   0.595097  -1.153 0.248982
## race3:estrogen_status1       1.820519   0.981128   1.856 0.063520 .
## race2:progesterone_status1  -0.050826   0.451620  -0.113 0.910394
## race3:progesterone_status1  -2.784800   0.946246  -2.943 0.003251 **
## race2:reginol_node_positive -0.121505   0.064782  -1.876 0.060712 .
## race3:reginol_node_positive -0.153094   0.089011  -1.720 0.085442 .
## race2:ln_tumor               0.776264   0.361491   2.147 0.031762 *
## race3:ln_tumor              -1.194039   0.762242  -1.566 0.117236
## race2:sqrt_examined         -0.085330   0.168637  -0.506 0.612859
## race3:sqrt_examined         -0.162959   0.213679  -0.763 0.445681
## race2:age                    0.008341   0.019360   0.431 0.666599
## race3:age                    0.035269   0.024348   1.449 0.147466
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3444.7  on 4023  degrees of freedom
## Residual deviance: 2898.0  on 3965  degrees of freedom
## AIC: 3016
##
## Number of Fisher Scoring iterations: 6
```

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

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

```
##     reginol_node_positive + ln_tumor + sqrt_examined + age +
##     race * (marital_status + progesterone_status + ln_tumor),
##     family = binomial, data = bc_ref)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  1.882124   0.611049   3.080 0.002069 **
## race2                       -1.131478   0.754804  -1.499 0.133864
## race3                        3.214161   1.374949   2.338 0.019405 *
## marital_status2             -0.186741   0.151851  -1.230 0.218785
## marital_status3              0.086222   0.158014   0.546 0.585299
## marital_status4             -0.076745   0.219242  -0.350 0.726303
## marital_status5             -0.780653   0.435332  -1.793 0.072935 .
## t_stage2                    -0.338441   0.160266  -2.112 0.034708 *
## t_stage3                    -0.375302   0.269563  -1.392 0.163844
## t_stage4                    -0.973119   0.315838  -3.081 0.002063 **
## n_stage2                    -0.495290   0.130427  -3.797 0.000146 ***
## n_stage3                    -0.664599   0.239253  -2.778 0.005473 **
## differentiate2               0.405297   0.105478   3.842 0.000122 ***
## differentiate3               0.947648   0.194613   4.869 1.12e-06 ***
## differentiate4              -0.997840   0.537651  -1.856 0.063464 .
## a_stage1                     0.057824   0.268010   0.216 0.829180
## estrogen_status1             0.768781   0.180131   4.268 1.97e-05 ***
## progesterone_status1         0.684689   0.134874   5.077 3.84e-07 ***
## reginol_node_positive       -0.075918   0.015092  -5.030 4.89e-07 ***
## ln_tumor                    -0.125372   0.146813  -0.854 0.393129
## sqrt_examined                0.258396   0.049886   5.180 2.22e-07 ***
## age                         -0.024851   0.005663  -4.389 1.14e-05 ***
## race2:marital_status2       -0.128238   0.518941  -0.247 0.804820
## race3:marital_status2       -0.373279   0.702629  -0.531 0.595238
## race2:marital_status3       -1.109358   0.385796  -2.876 0.004034 **
## race3:marital_status3       -1.349766   0.576548  -2.341 0.019226 *
## race2:marital_status4       -0.427490   0.551410  -0.775 0.438182
## race3:marital_status4       -1.600358   0.671688  -2.383 0.017191 *
## race2:marital_status5       -0.542786   1.006774  -0.539 0.589795
## race3:marital_status5       -0.911159   1.284113  -0.710 0.477975
## race2:progesterone_status1  -0.516652   0.357042  -1.447 0.147887
## race3:progesterone_status1  -1.646000   0.561267  -2.933 0.003361 **
## race2:ln_tumor               0.436542   0.213238   2.047 0.040638 *
## race3:ln_tumor              -0.336268   0.352855  -0.953 0.340594
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3444.7  on 4023  degrees of freedom
## Residual deviance: 2922.3  on 3990  degrees of freedom
## AIC: 2990.3
##
## Number of Fisher Scoring iterations: 5
```

- From the SRL for single race variable, we can see that the race group of black has a P-value of 6.95*10^(-6) and other race group has a P-value of 0.0257 both smaller than 0.05, so there are statistically significant difference between that to the reference of group white.

- From the MRL model without interaction terms, we can see that the AIC is around 2997.4 which is smaller than the original full model, when for both race2 and race3 we have a P-value smaller than 0.05 indicating significance, so we considering adding interaction terms.
- From the MRL model with race interaction terms added, we can see that the AIC is around 3016 which is getting larger, so we may have too many unnecessary interaction terms, and from the P-values we can see most the interaction covariates are insignificant except some terms in: race:marital_status, race:progesterone_status, race:ln_tumor.
- Therefore, we have the 3rd MRL model with race interaction terms selected, and we get a model with AIC of 2990.3 dropped sharply and the interaction term is included.

## Partial Test

### Partial Test for binary Y

```
# Our 1st global model for the predicts without colinearity and not normal predicts
model_global_bin = glm(status ~ race + marital_status + t_stage + n_stage +
                       differentiate + a_stage +estrogen_status + progesterone_status + ln_tumor + sqrt_

summary(model_global_bin)
```

```
##
## Call:
## glm(formula = status ~ race + marital_status + t_stage + n_stage +
##     differentiate + a_stage + estrogen_status + progesterone_status +
##     ln_tumor + sqrt_examined + age, family = binomial, data = newbc)
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)              2.025431   0.583766   3.470 0.000521 ***
## race2                   -0.490040   0.161448  -3.035 0.002403 **
## race3                    0.412102   0.201264   2.048 0.040602 *
## marital_status2         -0.213770   0.140919  -1.517 0.129273
## marital_status3         -0.149790   0.134179  -1.116 0.264275
## marital_status4         -0.220417   0.191076  -1.154 0.248681
## marital_status5         -0.936900   0.365551  -2.563 0.010378 *
## t_stage2                -0.355769   0.158618  -2.243 0.024902 *
## t_stage3                -0.450653   0.266574  -1.691 0.090925 .
## t_stage4                -1.044577   0.312476  -3.343 0.000829 ***
## n_stage2                -0.733552   0.117910  -6.221 4.93e-10 ***
## n_stage3                -1.575446   0.141592 -11.127  < 2e-16 ***
## differentiate2           0.375295   0.104473   3.592 0.000328 ***
## differentiate3           0.901142   0.192252   4.687 2.77e-06 ***
## differentiate4          -0.960780   0.535816  -1.793 0.072955 .
## a_stage1                 0.002227   0.262767   0.008 0.993239
## estrogen_status1         0.730261   0.176959   4.127 3.68e-05 ***
## progesterone_status1     0.573254   0.127304   4.503 6.70e-06 ***
## ln_tumor                -0.065530   0.138852  -0.472 0.636966
## sqrt_examined            0.180931   0.046777   3.868 0.000110 ***
## age                     -0.024590   0.005591  -4.398 1.09e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3444.7  on 4023  degrees of freedom
## Residual deviance: 2978.6  on 4003  degrees of freedom
## AIC: 3020.6
##
## Number of Fisher Scoring iterations: 5
```

From the summary of the global model, we can see that from the P-value of a_stage, ln_tumor, marital_status(2, 3, 4) have the p-value above 0.05 which is not significant. While differentiate has a 0.0729 p-value, and t_stage3 has a p-value of 0.0909, we may consider keep them. The covariates like t_stage, n_stage, race, estrogen_status, progesterone_status, sqrt_examined, age have a P-value smaller than 0.05 indicating significance. Therefore, our 1st partial test may include the significant covariates.

```
# 1st binary partial test with all significant covariates
# We will still keep marital_status because one group in this variable is significant
binmodel_partial_1 = glm(status ~ race + marital_status + t_stage + n_stage + differentiate + estrogen_

summary(binmodel_partial_1)
```

```
##
## Call:
## glm(formula = status ~ race + marital_status + t_stage + n_stage +
##     differentiate + estrogen_status + progesterone_status + sqrt_examined +
##     age, family = binomial, data = newbc)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           1.851463   0.377142   4.909 9.15e-07 ***
## race2                -0.488627   0.161358  -3.028 0.002460 **
## race3                 0.411208   0.201230   2.043 0.041006 *
## marital_status2      -0.215703   0.140825  -1.532 0.125594
## marital_status3      -0.151385   0.134089  -1.129 0.258903
## marital_status4      -0.221820   0.191055  -1.161 0.245631
## marital_status5      -0.938507   0.365357  -2.569 0.010207 *
## t_stage2             -0.408664   0.112638  -3.628 0.000285 ***
## t_stage3             -0.555367   0.148281  -3.745 0.000180 ***
## t_stage4             -1.125881   0.244474  -4.605 4.12e-06 ***
## n_stage2             -0.737853   0.117555  -6.277 3.46e-10 ***
## n_stage3             -1.581593   0.137038 -11.541  < 2e-16 ***
## differentiate2        0.377070   0.104385   3.612 0.000303 ***
## differentiate3        0.904754   0.191998   4.712 2.45e-06 ***
## differentiate4       -0.961568   0.535510  -1.796 0.072557 .
## estrogen_status1      0.730173   0.176789   4.130 3.62e-05 ***
## progesterone_status1  0.573495   0.127268   4.506 6.60e-06 ***
## sqrt_examined         0.181071   0.046740   3.874 0.000107 ***
## age                  -0.024489   0.005585  -4.385 1.16e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3444.7  on 4023  degrees of freedom
```

```
## Residual deviance: 2978.9  on 4005  degrees of freedom
## AIC: 3016.9
##
## Number of Fisher Scoring iterations: 5
```

```
# 2nd binary partial test with all in-significant covariates
binmodel_partial_2 = glm(status ~ a_stage + ln_tumor, family = binomial, newbc)

summary(binmodel_partial_2)
```

```
##
## Call:
## glm(formula = status ~ a_stage + ln_tumor, family = binomial,
##     data = newbc)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.53355    0.34110   7.428 1.11e-13 ***
## a_stage1     1.02938    0.22401   4.595 4.32e-06 ***
## ln_tumor    -0.55245    0.07032  -7.856 3.97e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3444.7  on 4023  degrees of freedom
## Residual deviance: 3351.7  on 4021  degrees of freedom
## AIC: 3357.7
##
## Number of Fisher Scoring iterations: 4
```

```
# 3rd binary partial test with some significant in 2nd test added
binmodel_partial_3 = glm(status ~ marital_status + t_stage + n_stage + differentiate + estrogen_status +

summary(binmodel_partial_3)
```

```
##
## Call:
## glm(formula = status ~ marital_status + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + sqrt_examined + age +
##     a_stage + ln_tumor, family = binomial, data = newbc)
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.985848   0.582219   3.411 0.000648 ***
## marital_status2     -0.247509   0.140237  -1.765 0.077575 .
## marital_status3     -0.231808   0.131581  -1.762 0.078119 .
## marital_status4     -0.281274   0.189215  -1.487 0.137138
## marital_status5     -0.986920   0.363073  -2.718 0.006563 **
## t_stage2            -0.346774   0.158600  -2.186 0.028781 *
## t_stage3            -0.450100   0.267133  -1.685 0.092003 .
## t_stage4            -1.056179   0.311620  -3.389 0.000701 ***
## n_stage2            -0.736774   0.117479  -6.272 3.57e-10 ***
```

```
## n_stage3             -1.584207   0.141279 -11.213  < 2e-16 ***
## differentiate2        0.390591   0.104008   3.755 0.000173 ***
## differentiate3        0.920784   0.191675   4.804 1.56e-06 ***
## differentiate4       -0.993288   0.534044  -1.860 0.062894 .
## estrogen_status1      0.731365   0.176189   4.151 3.31e-05 ***
## progesterone_status1  0.573479   0.126984   4.516 6.30e-06 ***
## sqrt_examined         0.183193   0.046609   3.930 8.48e-05 ***
## age                  -0.024564   0.005550  -4.426 9.60e-06 ***
## a_stage1             -0.004192   0.262185  -0.016 0.987243
## ln_tumor             -0.056204   0.139460  -0.403 0.686939
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3444.7  on 4023  degrees of freedom
## Residual deviance: 2993.1  on 4005  degrees of freedom
## AIC: 3031.1
##
## Number of Fisher Scoring iterations: 5
```

```r
# 4rd binary partial test with some selected interaction terms
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)

summary(model_inter)
```

```
##
## Call:
## glm(formula = status ~ race + marital_status + t_stage + n_stage +
##     differentiate + a_stage + estrogen_status + progesterone_status +
##     reginol_node_positive + ln_tumor + sqrt_examined + age +
##     race * (marital_status + progesterone_status + ln_tumor),
##     family = binomial, data = bc_ref)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  1.882124   0.611049   3.080 0.002069 **
## race2                       -1.131478   0.754804  -1.499 0.133864
## race3                        3.214161   1.374949   2.338 0.019405 *
## marital_status2             -0.186741   0.151851  -1.230 0.218785
## marital_status3              0.086222   0.158014   0.546 0.585299
## marital_status4             -0.076745   0.219242  -0.350 0.726303
## marital_status5             -0.780653   0.435332  -1.793 0.072935 .
## t_stage2                    -0.338441   0.160266  -2.112 0.034708 *
## t_stage3                    -0.375302   0.269563  -1.392 0.163844
## t_stage4                    -0.973119   0.315838  -3.081 0.002063 **
## n_stage2                    -0.495290   0.130427  -3.797 0.000146 ***
## n_stage3                    -0.664599   0.239253  -2.778 0.005473 **
## differentiate2               0.405297   0.105478   3.842 0.000122 ***
## differentiate3               0.947648   0.194613   4.869 1.12e-06 ***
## differentiate4              -0.997840   0.537651  -1.856 0.063464 .
```

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

- Our 1st partial test only include the significant predicts on global test, and then we also did a second partial test includes all the covariates that are non significant, and we can see the 1st partial model has an AIC at 3016.9. While the second model with all the non signifcant covraites only get an AIC at 3357.7, which is higher than the first model, so we will definitely not use the model 2 combination. However, we do discover some covariates are significant on the second model like a_stage, and ln_tumor, so we include these 2 again into the 1st model to get our 3rd model. However, our 3rd model has a higher AIC at 3031.1 than model 1, so we will continue to keep with model 1. Comparing Model 1 with the model 4 with some interaction terms included in race, we get a smaller AIC in model 4. Therefore, we will choose model 4 in our next step.

## Step-wise: forward/backward/AiC

### Step-wise for binary Y

```
binglobal_backward = stepAIC(model_global_bin, direction = "backward")
```

```
## Start:  AIC=3020.64
## status ~ race + marital_status + t_stage + n_stage + differentiate +
##     a_stage + estrogen_status + progesterone_status + ln_tumor +
##     sqrt_examined + age
##
```

```
##                        Df Deviance    AIC
## - a_stage               1   2978.6 3018.6
## - ln_tumor              1   2978.9 3018.9
## <none>                      2978.6 3020.6
## - marital_status        4   2987.5 3021.5
## - t_stage               3   2990.2 3026.2
## - race                  2   2993.1 3031.1
## - sqrt_examined         1   2993.7 3033.7
## - estrogen_status       1   2995.5 3035.5
## - progesterone_status   1   2997.9 3037.9
## - age                   1   2998.3 3038.3
## - differentiate         3   3011.0 3047.0
## - n_stage               2   3107.3 3145.3
##
## Step:  AIC=3018.64
## status ~ race + marital_status + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + ln_tumor + sqrt_examined +
##     age
##
##                        Df Deviance    AIC
## - ln_tumor              1   2978.9 3016.9
## <none>                      2978.6 3018.6
## - marital_status        4   2987.5 3019.5
## - t_stage               3   2991.2 3025.2
## - race                  2   2993.1 3029.1
## - sqrt_examined         1   2993.7 3031.7
## - estrogen_status       1   2995.5 3033.5
## - progesterone_status   1   2997.9 3035.9
## - age                   1   2998.3 3036.3
## - differentiate         3   3011.0 3045.0
## - n_stage               2   3113.8 3149.8
##
## Step:  AIC=3016.87
## status ~ race + marital_status + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + sqrt_examined + age
##
##                        Df Deviance    AIC
## <none>                      2978.9 3016.9
## - marital_status        4   2987.8 3017.8
## - race                  2   2993.2 3027.2
## - sqrt_examined         1   2994.0 3030.0
## - estrogen_status       1   2995.7 3031.7
## - progesterone_status   1   2998.1 3034.1
## - age                   1   2998.4 3034.4
## - t_stage               3   3008.6 3040.6
## - differentiate         3   3011.6 3043.6
## - n_stage               2   3116.4 3150.4
```

```
bin1_backward = stepAIC(binmodel_partial_1, direction = "backward")
```

```
## Start:  AIC=3016.87
## status ~ race + marital_status + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + sqrt_examined + age
##
```

```
##                       Df Deviance    AIC
## <none>                     2978.9 3016.9
## - marital_status       4   2987.8 3017.8
## - race                 2   2993.2 3027.2
## - sqrt_examined        1   2994.0 3030.0
## - estrogen_status      1   2995.7 3031.7
## - progesterone_status  1   2998.1 3034.1
## - age                  1   2998.4 3034.4
## - t_stage              3   3008.6 3040.6
## - differentiate        3   3011.6 3043.6
## - n_stage              2   3116.4 3150.4
```

```
bin2_backward = stepAIC(binmodel_partial_2, direction = "backward")
```

```
## Start:  AIC=3357.74
## status ~ a_stage + ln_tumor
##
##            Df Deviance    AIC
## <none>          3351.7 3357.7
## - a_stage   1   3370.9 3374.9
## - ln_tumor  1   3415.7 3419.7
```

```
bin3_backward = stepAIC(binmodel_partial_3, direction = "backward")
```

```
## Start:  AIC=3031.09
## status ~ marital_status + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + sqrt_examined + age +
##     a_stage + ln_tumor
##
##                       Df Deviance    AIC
## - a_stage              1   2993.1 3029.1
## - ln_tumor             1   2993.2 3029.2
## <none>                     2993.1 3031.1
## - marital_status       4   3005.1 3035.1
## - t_stage              3   3004.8 3036.8
## - sqrt_examined        1   3008.6 3044.6
## - estrogen_status      1   3010.1 3046.1
## - progesterone_status  1   3012.4 3048.4
## - age                  1   3013.0 3049.0
## - differentiate        3   3027.6 3059.6
## - n_stage              2   3123.8 3157.8
##
## Step:  AIC=3029.09
## status ~ marital_status + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + sqrt_examined + age +
##     ln_tumor
##
##                       Df Deviance    AIC
## - ln_tumor             1   2993.2 3027.2
## <none>                     2993.1 3029.1
## - marital_status       4   3005.1 3033.1
## - t_stage              3   3005.8 3035.8
## - sqrt_examined        1   3008.6 3042.6
```

```
## - estrogen_status      1   3010.1 3044.1
## - progesterone_status  1   3012.4 3046.4
## - age                  1   3013.1 3047.1
## - differentiate        3   3027.6 3057.6
## - n_stage              2   3130.2 3162.2
##
## Step:  AIC=3027.25
## status ~ marital_status + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + sqrt_examined + age
##
##                       Df Deviance    AIC
## <none>                     2993.2 3027.3
## - marital_status       4   3005.3 3031.3
## - sqrt_examined        1   3008.8 3040.8
## - estrogen_status      1   3010.3 3042.3
## - progesterone_status  1   3012.6 3044.6
## - age                  1   3013.1 3045.1
## - t_stage              3   3022.2 3050.2
## - differentiate        3   3028.1 3056.1
## - n_stage              2   3132.5 3162.5
```

```
bin4_backward = stepAIC(model_inter, direction = "backward")
```

```
## Start:  AIC=2990.34
## 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)
##
##                             Df Deviance    AIC
## - a_stage                    1   2922.4 2988.4
## <none>                           2922.3 2990.3
## - race:marital_status        8   2939.2 2991.2
## - race:ln_tumor              2   2927.7 2991.7
## - t_stage                    3   2932.8 2994.8
## - race:progesterone_status   2   2934.0 2998.0
## - n_stage                    2   2937.3 3001.3
## - estrogen_status            1   2940.3 3006.3
## - age                        1   2941.9 3007.9
## - reginol_node_positive      1   2947.9 3013.9
## - sqrt_examined              1   2949.6 3015.6
## - differentiate              3   2957.8 3019.8
##
## Step:  AIC=2988.38
## status ~ race + marital_status + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + reginol_node_positive +
##     ln_tumor + sqrt_examined + age + race:marital_status + race:progesterone_status +
##     race:ln_tumor
##
##                             Df Deviance    AIC
## <none>                           2922.4 2988.4
## - race:marital_status        8   2939.2 2989.2
## - race:ln_tumor              2   2927.7 2989.7
## - t_stage                    3   2934.3 2994.3
```

```
## - race:progesterone_status  2   2934.0 2996.0
## - n_stage                   2   2937.5 2999.5
## - estrogen_status           1   2940.5 3004.5
## - age                       1   2941.9 3005.9
## - reginol_node_positive     1   2947.9 3011.9
## - sqrt_examined             1   2949.7 3013.7
## - differentiate             3   2957.8 3017.8
```

From the backward elimination, the best model with lowest AIC of 2988.32 is: status ~ race + marital_status + t_stage + n_stage + differentiate + estrogen_status + progesterone_status + reginol_node_positive + ln_tumor + sqrt_examined + age + race:marital_status + race:progesterone_status + race:ln_tumor.

```
binglobal_backward = stepAIC(model_global_bin, direction = "forward")
```

```
## Start:  AIC=3020.64
## status ~ race + marital_status + t_stage + n_stage + differentiate +
##     a_stage + estrogen_status + progesterone_status + ln_tumor +
##     sqrt_examined + age
```

```
bin1_backward = stepAIC(binmodel_partial_1, direction = "forward")
```

```
## Start:  AIC=3016.87
## status ~ race + marital_status + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + sqrt_examined + age
```

```
bin2_backward = stepAIC(binmodel_partial_2, direction = "forward")
```

```
## Start:  AIC=3357.74
## status ~ a_stage + ln_tumor
```

```
bin3_backward = stepAIC(binmodel_partial_3, direction = "forward")
```

```
## Start:  AIC=3031.09
## status ~ marital_status + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + sqrt_examined + age +
##     a_stage + ln_tumor
```

```
bin4_backward = stepAIC(model_inter, direction = "forward")
```

```
## Start:  AIC=2990.34
## 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)
```

From the forward selection, the best model with lowest AIC of 2990.34 status ~ race + marital_status + t_stage + n_stage + differentiate + estrogen_status + progesterone_status + reginol_node_positive + ln_tumor + sqrt_examined + age + race:marital_status + race:progesterone_status + race:ln_tumor. Backward and forward elimination selection both produce the same model.

```r
binglobal_backward = stepAIC(model_global_bin, direction = "both")
```

```
## Start:  AIC=3020.64
## status ~ race + marital_status + t_stage + n_stage + differentiate +
##     a_stage + estrogen_status + progesterone_status + ln_tumor +
##     sqrt_examined + age
##
##                       Df Deviance    AIC
## - a_stage              1   2978.6 3018.6
## - ln_tumor             1   2978.9 3018.9
## <none>                     2978.6 3020.6
## - marital_status       4   2987.5 3021.5
## - t_stage              3   2990.2 3026.2
## - race                 2   2993.1 3031.1
## - sqrt_examined        1   2993.7 3033.7
## - estrogen_status      1   2995.5 3035.5
## - progesterone_status  1   2997.9 3037.9
## - age                  1   2998.3 3038.3
## - differentiate        3   3011.0 3047.0
## - n_stage              2   3107.3 3145.3
##
## Step:  AIC=3018.64
## status ~ race + marital_status + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + ln_tumor + sqrt_examined +
##     age
##
##                       Df Deviance    AIC
## - ln_tumor             1   2978.9 3016.9
## <none>                     2978.6 3018.6
## - marital_status       4   2987.5 3019.5
## + a_stage              1   2978.6 3020.6
## - t_stage              3   2991.2 3025.2
## - race                 2   2993.1 3029.1
## - sqrt_examined        1   2993.7 3031.7
## - estrogen_status      1   2995.5 3033.5
## - progesterone_status  1   2997.9 3035.9
## - age                  1   2998.3 3036.3
## - differentiate        3   3011.0 3045.0
## - n_stage              2   3113.8 3149.8
##
## Step:  AIC=3016.87
## status ~ race + marital_status + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + sqrt_examined + age
##
##                       Df Deviance    AIC
## <none>                     2978.9 3016.9
## - marital_status       4   2987.8 3017.8
## + ln_tumor             1   2978.6 3018.6
## + a_stage              1   2978.9 3018.9
## - race                 2   2993.2 3027.2
## - sqrt_examined        1   2994.0 3030.0
## - estrogen_status      1   2995.7 3031.7
## - progesterone_status  1   2998.1 3034.1
```

```
## - age                  1   2998.4 3034.4
## - t_stage              3   3008.6 3040.6
## - differentiate        3   3011.6 3043.6
## - n_stage              2   3116.4 3150.4
```

```
bin1_backward = stepAIC(binmodel_partial_1, direction = "both")
```

```
## Start:  AIC=3016.87
## status ~ race + marital_status + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + sqrt_examined + age
##
##                       Df Deviance    AIC
## <none>                     2978.9 3016.9
## - marital_status       4   2987.8 3017.8
## - race                 2   2993.2 3027.2
## - sqrt_examined        1   2994.0 3030.0
## - estrogen_status      1   2995.7 3031.7
## - progesterone_status  1   2998.1 3034.1
## - age                  1   2998.4 3034.4
## - t_stage              3   3008.6 3040.6
## - differentiate        3   3011.6 3043.6
## - n_stage              2   3116.4 3150.4
```

```
bin2_backward = stepAIC(binmodel_partial_2, direction = "both")
```

```
## Start:  AIC=3357.74
## status ~ a_stage + ln_tumor
##
##            Df Deviance    AIC
## <none>          3351.7 3357.7
## - a_stage   1   3370.9 3374.9
## - ln_tumor  1   3415.7 3419.7
```

```
bin3_backward = stepAIC(binmodel_partial_3, direction = "both")
```

```
## Start:  AIC=3031.09
## status ~ marital_status + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + sqrt_examined + age +
##     a_stage + ln_tumor
##
##                       Df Deviance    AIC
## - a_stage              1   2993.1 3029.1
## - ln_tumor             1   2993.2 3029.2
## <none>                     2993.1 3031.1
## - marital_status       4   3005.1 3035.1
## - t_stage              3   3004.8 3036.8
## - sqrt_examined        1   3008.6 3044.6
## - estrogen_status      1   3010.1 3046.1
## - progesterone_status  1   3012.4 3048.4
## - age                  1   3013.0 3049.0
## - differentiate        3   3027.6 3059.6
## - n_stage              2   3123.8 3157.8
```

```
##
## Step:  AIC=3029.09
## status ~ marital_status + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + sqrt_examined + age +
##     ln_tumor
##
##                       Df Deviance    AIC
## - ln_tumor             1   2993.2 3027.2
## <none>                     2993.1 3029.1
## + a_stage              1   2993.1 3031.1
## - marital_status       4   3005.1 3033.1
## - t_stage              3   3005.8 3035.8
## - sqrt_examined        1   3008.6 3042.6
## - estrogen_status      1   3010.1 3044.1
## - progesterone_status  1   3012.4 3046.4
## - age                  1   3013.1 3047.1
## - differentiate        3   3027.6 3057.6
## - n_stage              2   3130.2 3162.2
##
## Step:  AIC=3027.25
## status ~ marital_status + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + sqrt_examined + age
##
##                       Df Deviance    AIC
## <none>                     2993.2 3027.3
## + ln_tumor             1   2993.1 3029.1
## + a_stage              1   2993.2 3029.3
## - marital_status       4   3005.3 3031.3
## - sqrt_examined        1   3008.8 3040.8
## - estrogen_status      1   3010.3 3042.3
## - progesterone_status  1   3012.6 3044.6
## - age                  1   3013.1 3045.1
## - t_stage              3   3022.2 3050.2
## - differentiate        3   3028.1 3056.1
## - n_stage              2   3132.5 3162.5
```

```
bin4_backward = stepAIC(model_inter, direction = "both")
```

```
## Start:  AIC=2990.34
## 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)
##
##                              Df Deviance    AIC
## - a_stage                     1   2922.4 2988.4
## <none>                            2922.3 2990.3
## - race:marital_status         8   2939.2 2991.2
## - race:ln_tumor               2   2927.7 2991.7
## - t_stage                     3   2932.8 2994.8
## - race:progesterone_status    2   2934.0 2998.0
## - n_stage                     2   2937.3 3001.3
## - estrogen_status             1   2940.3 3006.3
## - age                         1   2941.9 3007.9
```

```
## - reginol_node_positive     1   2947.9 3013.9
## - sqrt_examined             1   2949.6 3015.6
## - differentiate             3   2957.8 3019.8
##
## Step:  AIC=2988.38
## status ~ race + marital_status + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + reginol_node_positive +
##     ln_tumor + sqrt_examined + age + race:marital_status + race:progesterone_status +
##     race:ln_tumor
##
##                            Df Deviance    AIC
## <none>                          2922.4 2988.4
## - race:marital_status       8   2939.2 2989.2
## - race:ln_tumor             2   2927.7 2989.7
## + a_stage                   1   2922.3 2990.3
## - t_stage                   3   2934.3 2994.3
## - race:progesterone_status  2   2934.0 2996.0
## - n_stage                   2   2937.5 2999.5
## - estrogen_status           1   2940.5 3004.5
## - age                       1   2941.9 3005.9
## - reginol_node_positive     1   2947.9 3011.9
## - sqrt_examined             1   2949.7 3013.7
## - differentiate             3   2957.8 3017.8
```

- From the stepwise regression, we have the lowest AIC of 3016.98 which is the model of status ~ race + marital_status + t_stage + n_stage + differentiate + estrogen_status + progesterone_status + reginol_node_positive + ln_tumor + sqrt_examined + age + race:marital_status + race:progesterone_status + race:ln_tumor. The model is the same compared to the one we picked from the backward elimination and forward elimination. We will go with this model for further statistical testing.

## ANOVA

```
model2 <- glm(status ~ race + marital_status + t_stage + n_stage + differentiate + estrogen_status + pr
anova_result1 <- anova(model2)
print(anova_result1)
```

### anova for binary Y

```
## Analysis of Deviance Table
##
## Model: gaussian, link: identity
##
## Response: status
##
## Terms added sequentially (first to last)
##
##
##                     Df Deviance Resid. Df Resid. Dev
## NULL                                 4023     521.70
```

```
## race                 2   3.6262      4021     518.08
## marital_status       4   2.9196      4017     515.16
## t_stage              3  13.1374      4014     502.02
## n_stage              2  25.1703      4012     476.85
## differentiate        3   6.8813      4009     469.97
## estrogen_status      1   8.9916      4008     460.98
## progesterone_status  1   2.9595      4007     458.02
## sqrt_examined        1   1.8003      4006     456.22
## age                  1   2.1022      4005     454.11
```

```
summary(model2)
```

```
##
## Call:
## glm(formula = status ~ race + marital_status + t_stage + n_stage +
##     differentiate + estrogen_status + progesterone_status + sqrt_examined +
##     age, data = newbc)
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           0.7708680  0.0439145  17.554  < 2e-16 ***
## race2                -0.0681853  0.0209930  -3.248 0.001172 **
## race3                 0.0416639  0.0198151   2.103 0.035560 *
## marital_status2      -0.0259111  0.0167149  -1.550 0.121178
## marital_status3      -0.0175357  0.0153627  -1.141 0.253752
## marital_status4      -0.0321326  0.0235225  -1.366 0.172004
## marital_status5      -0.1345247  0.0508574  -2.645 0.008198 **
## t_stage2             -0.0396395  0.0118704  -3.339 0.000847 ***
## t_stage3             -0.0603490  0.0174234  -3.464 0.000538 ***
## t_stage4             -0.1748054  0.0350531  -4.987 6.40e-07 ***
## n_stage2             -0.0837902  0.0139153  -6.021 1.88e-09 ***
## n_stage3             -0.2421576  0.0183695 -13.183  < 2e-16 ***
## differentiate2        0.0509827  0.0127519   3.998 6.50e-05 ***
## differentiate3        0.0895198  0.0183190   4.887 1.07e-06 ***
## differentiate4       -0.1824586  0.0781803  -2.334 0.019654 *
## estrogen_status1      0.1408815  0.0252403   5.582 2.54e-08 ***
## progesterone_status1  0.0779114  0.0164916   4.724 2.39e-06 ***
## sqrt_examined         0.0191703  0.0050357   3.807 0.000143 ***
## age                  -0.0026658  0.0006191  -4.306 1.70e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1133866)
##
##     Null deviance: 521.70  on 4023  degrees of freedom
## Residual deviance: 454.11  on 4005  degrees of freedom
## AIC: 2680.5
##
## Number of Fisher Scoring iterations: 2
```

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

anova(null_model, model2, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: status ~ 1
## Model 2: status ~ race + marital_status + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + sqrt_examined + age
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      4023     3444.7
## 2      4005      454.1 18   2990.6 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With p significant p value, we have enough evidence to show that the full model is better, which is status ~ race + marital_status + t_stage + n_stage + differentiate + estrogen_status + progesterone_status + sqrt_examined + age

## Criterion procedures

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

```
## Start:  AIC=2990.34
## 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)
##
##                             Df Deviance    AIC
## - a_stage                    1   2922.4 2988.4
## <none>                           2922.3 2990.3
## - race:marital_status        8   2939.2 2991.2
## - race:ln_tumor              2   2927.7 2991.7
## - t_stage                    3   2932.8 2994.8
## - race:progesterone_status   2   2934.0 2998.0
## - n_stage                    2   2937.3 3001.3
## - estrogen_status            1   2940.3 3006.3
## - age                        1   2941.9 3007.9
## - reginol_node_positive      1   2947.9 3013.9
## - sqrt_examined              1   2949.6 3015.6
## - differentiate              3   2957.8 3019.8
##
## Step:  AIC=2988.38
## status ~ race + marital_status + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + reginol_node_positive +
##     ln_tumor + sqrt_examined + age + race:marital_status + race:progesterone_status +
##     race:ln_tumor
##
##                             Df Deviance    AIC
## <none>                           2922.4 2988.4
## - race:marital_status        8   2939.2 2989.2
## - race:ln_tumor              2   2927.7 2989.7
## + a_stage                    1   2922.3 2990.3
## - t_stage                    3   2934.3 2994.3
```

```
## - race:progesterone_status  2   2934.0 2996.0
## - n_stage                   2   2937.5 2999.5
## - estrogen_status           1   2940.5 3004.5
## - age                       1   2941.9 3005.9
## - reginol_node_positive     1   2947.9 3011.9
## - sqrt_examined             1   2949.7 3013.7
## - differentiate             3   2957.8 3017.8
```

```
summary(aic_selected_bin_model)
```

```
##
## Call:
## glm(formula = status ~ race + marital_status + t_stage + n_stage +
##     differentiate + estrogen_status + progesterone_status + reginol_node_positive +
##     ln_tumor + sqrt_examined + age + race:marital_status + race:progesterone_status +
##     race:ln_tumor, family = binomial, data = bc_ref)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  1.935909   0.557872   3.470 0.000520 ***
## race2                       -1.132866   0.754835  -1.501 0.133404
## race3                        3.216611   1.374740   2.340 0.019294 *
## marital_status2             -0.187047   0.151846  -1.232 0.218018
## marital_status3              0.085354   0.157948   0.540 0.588926
## marital_status4             -0.076369   0.219281  -0.348 0.727637
## marital_status5             -0.780638   0.435255  -1.794 0.072890 .
## t_stage2                    -0.339271   0.160216  -2.118 0.034210 *
## t_stage3                    -0.376648   0.269485  -1.398 0.162215
## t_stage4                    -0.994213   0.300240  -3.311 0.000928 ***
## n_stage2                    -0.495892   0.130402  -3.803 0.000143 ***
## n_stage3                    -0.672536   0.236499  -2.844 0.004459 **
## differentiate2               0.404977   0.105469   3.840 0.000123 ***
## differentiate3               0.946644   0.194534   4.866 1.14e-06 ***
## differentiate4              -0.994404   0.537987  -1.848 0.064547 .
## estrogen_status1             0.770032   0.180046   4.277 1.90e-05 ***
## progesterone_status1         0.684038   0.134861   5.072 3.93e-07 ***
## reginol_node_positive       -0.075837   0.015094  -5.024 5.05e-07 ***
## ln_tumor                    -0.124688   0.146757  -0.850 0.395537
## sqrt_examined                0.258542   0.049878   5.184 2.18e-07 ***
## age                         -0.024823   0.005661  -4.385 1.16e-05 ***
## race2:marital_status2       -0.124157   0.518767  -0.239 0.810849
## race3:marital_status2       -0.375881   0.702191  -0.535 0.592444
## race2:marital_status3       -1.107998   0.385723  -2.873 0.004072 **
## race3:marital_status3       -1.348021   0.576449  -2.338 0.019362 *
## race2:marital_status4       -0.425603   0.551436  -0.772 0.440228
## race3:marital_status4       -1.600030   0.671850  -2.382 0.017241 *
## race2:marital_status5       -0.558575   1.004035  -0.556 0.577985
## race3:marital_status5       -0.909319   1.283963  -0.708 0.478813
## race2:progesterone_status1  -0.516300   0.357063  -1.446 0.148188
## race3:progesterone_status1  -1.642600   0.560876  -2.929 0.003405 **
## race2:ln_tumor               0.436696   0.213247   2.048 0.040576 *
## race3:ln_tumor              -0.337919   0.352801  -0.958 0.338154
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3444.7  on 4023  degrees of freedom
## Residual deviance: 2922.4  on 3991  degrees of freedom
## AIC: 2988.4
##
## Number of Fisher Scoring iterations: 5
```

```
bic_selected_bin_model <- stepAIC(model_inter, direction = "both", family = binomial, k = log(nrow(bc_r
```

```
## Start:  AIC=3204.54
## 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)
##
##                             Df Deviance    AIC
## - race:marital_status        8   2939.2 3155.0
## - t_stage                    3   2932.8 3190.2
## - race:ln_tumor              2   2927.7 3193.3
## - a_stage                    1   2922.4 3196.3
## - race:progesterone_status   2   2934.0 3199.6
## - n_stage                    2   2937.3 3202.9
## <none>                           2922.3 3204.5
## - estrogen_status            1   2940.3 3214.2
## - differentiate              3   2957.8 3215.2
## - age                        1   2941.9 3215.8
## - reginol_node_positive      1   2947.9 3221.8
## - sqrt_examined              1   2949.6 3223.5
##
## Step:  AIC=3155.01
## status ~ race + marital_status + t_stage + n_stage + differentiate +
##     a_stage + estrogen_status + progesterone_status + reginol_node_positive +
##     ln_tumor + sqrt_examined + age + race:progesterone_status +
##     race:ln_tumor
##
##                             Df Deviance    AIC
## - marital_status             4   2947.4 3130.1
## - t_stage                    3   2950.0 3140.9
## - race:ln_tumor              2   2943.9 3143.1
## - a_stage                    1   2939.2 3146.7
## - race:progesterone_status   2   2949.1 3148.3
## - n_stage                    2   2952.6 3151.8
## <none>                           2939.2 3155.0
## - differentiate              3   2973.2 3164.1
## - estrogen_status            1   2957.7 3165.2
## - age                        1   2958.4 3165.9
## - reginol_node_positive      1   2965.8 3173.3
## - sqrt_examined              1   2966.4 3173.9
## + race:marital_status        8   2922.3 3204.5
##
## Step:  AIC=3130.05
## status ~ race + t_stage + n_stage + differentiate + a_stage +
```

```
##     estrogen_status + progesterone_status + reginol_node_positive +
##     ln_tumor + sqrt_examined + age + race:progesterone_status +
##     race:ln_tumor
##
##                             Df Deviance    AIC
## - t_stage                    3   2958.4 3116.1
## - race:ln_tumor              2   2952.0 3118.0
## - a_stage                    1   2947.5 3121.8
## - race:progesterone_status   2   2957.0 3123.0
## - n_stage                    2   2960.7 3126.7
## <none>                           2947.4 3130.1
## - differentiate              3   2981.3 3139.0
## - estrogen_status            1   2965.6 3139.9
## - age                        1   2968.6 3142.9
## - reginol_node_positive      1   2975.2 3149.5
## - sqrt_examined              1   2975.3 3149.6
## + marital_status             4   2939.2 3155.0
##
## Step:  AIC=3116.1
## status ~ race + n_stage + differentiate + a_stage + estrogen_status +
##     progesterone_status + reginol_node_positive + ln_tumor +
##     sqrt_examined + age + race:progesterone_status + race:ln_tumor
##
##                             Df Deviance    AIC
## - race:ln_tumor              2   2963.3 3104.4
## - race:progesterone_status   2   2968.0 3109.1
## - a_stage                    1   2960.0 3109.4
## - n_stage                    2   2972.3 3113.4
## <none>                           2958.4 3116.1
## - estrogen_status            1   2977.0 3126.4
## - differentiate              3   2994.0 3126.8
## - age                        1   2980.4 3129.8
## + t_stage                    3   2947.4 3130.1
## - sqrt_examined              1   2986.1 3135.5
## - reginol_node_positive      1   2986.6 3136.0
## + marital_status             4   2950.0 3140.9
##
## Step:  AIC=3104.4
## status ~ race + n_stage + differentiate + a_stage + estrogen_status +
##     progesterone_status + reginol_node_positive + ln_tumor +
##     sqrt_examined + age + race:progesterone_status
##
##                             Df Deviance    AIC
## - race:progesterone_status   2   2972.7 3097.2
## - a_stage                    1   2964.9 3097.7
## - n_stage                    2   2977.2 3101.7
## <none>                           2963.3 3104.4
## - ln_tumor                   1   2979.9 3112.7
## - differentiate              3   2998.1 3114.3
## - estrogen_status            1   2982.0 3114.8
## + race:ln_tumor              2   2958.4 3116.1
## - age                        1   2985.0 3117.8
## + t_stage                    3   2952.0 3118.0
## - sqrt_examined              1   2990.6 3123.4
```

```
## - reginol_node_positive     1   2990.9 3123.7
## + marital_status            4   2955.1 3129.4
##
## Step:  AIC=3097.15
## status ~ race + n_stage + differentiate + a_stage + estrogen_status +
##     progesterone_status + reginol_node_positive + ln_tumor +
##     sqrt_examined + age
##
##                             Df Deviance    AIC
## - a_stage                    1   2974.2 3090.4
## - n_stage                    2   2987.1 3095.0
## <none>                           2972.7 3097.2
## - race                       2   2991.8 3099.7
## + race:progesterone_status   2   2963.3 3104.4
## - estrogen_status            1   2989.9 3106.1
## - ln_tumor                   1   2989.9 3106.1
## - differentiate              3   3007.4 3107.0
## + race:ln_tumor              2   2968.0 3109.1
## - progesterone_status        1   2993.3 3109.5
## - age                        1   2993.4 3109.7
## + t_stage                    3   2961.2 3110.6
## - reginol_node_positive      1   2999.2 3115.4
## - sqrt_examined              1   2999.9 3116.1
## + marital_status             4   2964.7 3122.4
##
## Step:  AIC=3090.36
## status ~ race + n_stage + differentiate + estrogen_status + progesterone_status +
##     reginol_node_positive + ln_tumor + sqrt_examined + age
##
##                             Df Deviance    AIC
## - n_stage                    2   2989.3 3088.9
## <none>                           2974.2 3090.4
## - race                       2   2993.3 3092.9
## + a_stage                    1   2972.7 3097.2
## + race:progesterone_status   2   2964.9 3097.7
## - estrogen_status            1   2991.9 3099.8
## - differentiate              3   3008.5 3099.8
## - ln_tumor                   1   2992.1 3100.0
## + race:ln_tumor              2   2969.5 3102.3
## + t_stage                    3   2961.2 3102.3
## - progesterone_status        1   2994.5 3102.4
## - age                        1   2994.7 3102.6
## - reginol_node_positive      1   3000.4 3108.3
## - sqrt_examined              1   3001.7 3109.6
## + marital_status             4   2966.0 3115.4
##
## Step:  AIC=3088.9
## status ~ race + differentiate + estrogen_status + progesterone_status +
##     reginol_node_positive + ln_tumor + sqrt_examined + age
##
##                             Df Deviance    AIC
## <none>                           2989.3 3088.9
## + n_stage                    2   2974.2 3090.4
## - race                       2   3009.5 3092.5
```

```
## + a_stage                    1   2987.1 3095.0
## + race:progesterone_status   2   2979.6 3095.8
## - estrogen_status            1   3008.0 3099.3
## + t_stage                    3   2975.1 3099.6
## + race:ln_tumor              2   2984.6 3100.8
## - age                        1   3009.9 3101.2
## - progesterone_status        1   3010.3 3101.6
## - differentiate              3   3027.0 3101.7
## - ln_tumor                   1   3013.8 3105.1
## - sqrt_examined              1   3015.5 3106.8
## + marital_status             4   2981.4 3114.2
## - reginol_node_positive      1   3146.2 3237.5
```

```
summary(bic_selected_bin_model)
```

```
##
## Call:
## glm(formula = status ~ race + differentiate + 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)              2.451459   0.448634   5.464 4.65e-08 ***
## race2                   -0.595023   0.157272  -3.783 0.000155 ***
## race3                    0.440768   0.201932   2.183 0.029054 *
## differentiate2           0.417805   0.103868   4.022 5.76e-05 ***
## differentiate3           0.957010   0.190932   5.012 5.38e-07 ***
## differentiate4          -0.916981   0.506763  -1.809 0.070375 .
## estrogen_status1         0.763842   0.175498   4.352 1.35e-05 ***
## progesterone_status1     0.596058   0.126629   4.707 2.51e-06 ***
## reginol_node_positive -0.109570   0.008846 -12.386  < 2e-16 ***
## ln_tumor                -0.366801   0.075028  -4.889 1.01e-06 ***
## sqrt_examined            0.248487   0.048812   5.091 3.57e-07 ***
## age                     -0.024373   0.005417  -4.500 6.80e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3444.7  on 4023  degrees of freedom
## Residual deviance: 2989.3  on 4012  degrees of freedom
## AIC: 3013.3
##
## Number of Fisher Scoring iterations: 5
```

```
bic_selected_model <- stepAIC(bic_selected_bin_model, direction = "backward")
```

```
## Start:  AIC=3013.3
## status ~ race + differentiate + estrogen_status + progesterone_status +
##     reginol_node_positive + ln_tumor + sqrt_examined + age
##
##                          Df Deviance    AIC
```

```
## <none>                       2989.3 3013.3
## - race                    2   3009.5 3029.5
## - estrogen_status         1   3008.0 3030.0
## - age                     1   3009.9 3031.9
## - progesterone_status     1   3010.3 3032.3
## - ln_tumor                1   3013.8 3035.8
## - sqrt_examined           1   3015.5 3037.5
## - differentiate           3   3027.0 3045.0
## - reginol_node_positive   1   3146.2 3168.2
```

```
bic_selected_model <- stepAIC(bic_selected_bin_model, direction = "backward")
```

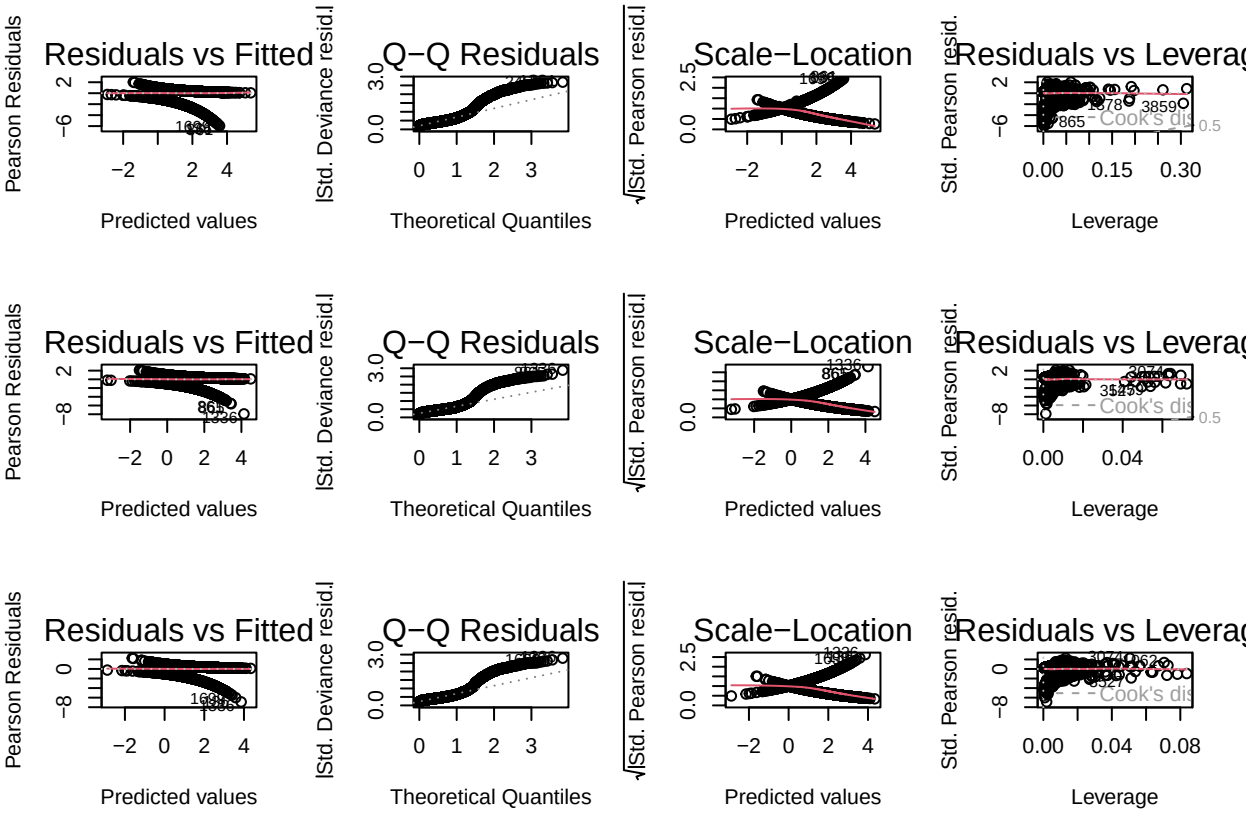
```
## Start:  AIC=3013.3
## status ~ race + differentiate + estrogen_status + progesterone_status +
##     reginol_node_positive + ln_tumor + sqrt_examined + age
##
##                          Df Deviance    AIC
## <none>                       2989.3 3013.3
## - race                    2   3009.5 3029.5
## - estrogen_status         1   3008.0 3030.0
## - age                     1   3009.9 3031.9
## - progesterone_status     1   3010.3 3032.3
## - ln_tumor                1   3013.8 3035.8
## - sqrt_examined           1   3015.5 3037.5
## - differentiate           3   3027.0 3045.0
## - reginol_node_positive   1   3146.2 3168.2
```

In the interaction model, the AIC selected model is status ~ race + marital_status + t_stage + n_stage + differentiate + estrogen_status + progesterone_status + reginol_node_positive + ln_tumor + sqrt_examined + age + race:marital_status + race:progesterone_status + race:ln_tumor.

The BIC selected model is status ~ race + differentiate + estrogen_status + progesterone_status + reginol_node_positive + ln_tumor + sqrt_examined + age.

**Plots to check the model assumptions:**

```
par(mfrow = c(3,4))
plot(aic_selected_bin_model)
plot(bic_selected_bin_model)
plot(binmodel_partial_1)
```




**VIF**

```
# Calculate VIF for the model selected by AIC
vif_result_aic <- vif(aic_selected_bin_model, type = 'predictor')
```

```
## Warning in vif.lm(aic_selected_bin_model, type = "predictor"): type = 'predictor' is available only i
##   type = 'terms' will be used
```

```
print(vif_result_aic)
```

```
##                                GVIF Df GVIF^(1/(2*Df))
## race                    1077.779523  2        5.729708
## marital_status             3.204153  4        1.156682
## t_stage                    3.798503  3        1.249114
## n_stage                    3.906627  2        1.405887
## differentiate              1.137198  3        1.021659
## estrogen_status            1.476209  1        1.214993
## progesterone_status        1.567471  1        1.251987
## reginol_node_positive      4.091601  1        2.022771
## ln_tumor                   4.114980  1        2.028541
## sqrt_examined              1.426654  1        1.194426
## age                        1.123538  1        1.059971
## race:marital_status       15.664545  8        1.187633
## race:progesterone_status  25.681757  2        2.251159
## race:ln_tumor            772.595375  2        5.272153
```

```r
# Calculate VIF for the model selected by BIC
vif_result_bic <- vif(bic_selected_bin_model)
print(vif_result_bic)
```

```
##                           GVIF Df GVIF^(1/(2*Df))
## race                  1.018540  2        1.004603
## differentiate         1.102015  3        1.016322
## estrogen_status       1.474559  1        1.214314
## progesterone_status   1.427838  1        1.194922
## reginol_node_positive 1.415218  1        1.189629
## ln_tumor              1.066185  1        1.032563
## sqrt_examined         1.389686  1        1.178849
## age                   1.043326  1        1.021433
```

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 multicolinearity (VIF)) 7. cross validation