Final Project

Shiying Wu, Maggie Hsu, Xun Sun

2024-12-19

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.1 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(survival)  
library(caret)

## Warning: 程序包'caret'是用R版本4.4.2 来建造的

## 载入需要的程序包：lattice  
##   
## 载入程序包：'caret'  
##   
## The following object is masked from 'package:survival':  
##   
## cluster  
##   
## The following object is masked from 'package:purrr':  
##   
## lift

#car package for vif  
library(car)

## Warning: 程序包'car'是用R版本4.4.2 来建造的

## 载入需要的程序包：carData

## Warning: 程序包'carData'是用R版本4.4.2 来建造的

##   
## 载入程序包：'car'  
##   
## The following object is masked from 'package:dplyr':  
##   
## recode  
##   
## The following object is masked from 'package:purrr':  
##   
## some

#MASS package for var selection  
library(MASS)

## Warning: 程序包'MASS'是用R版本4.4.2 来建造的

##   
## 载入程序包：'MASS'  
##   
## The following object is masked from 'package:dplyr':  
##   
## select

data <- read.csv("Project\_2\_data.csv")  
view(data)  
  
  
  
sampledf <- data |>  
 mutate(Status = ifelse(Status == "Dead", 1, 0))  
  
dead <- sampledf |>  
 filter(Status == 1)  
  
alive <- sampledf |>  
 filter(Status == 0) |>  
 sample\_n(nrow(dead))  
  
final<-bind\_rows(dead, alive) |>  
 dplyr::select(-Survival.Months)  
  
logit\_model <- glm(Status ~ ., data = final, family = binomial())  
  
summary(logit\_model)

##   
## Call:  
## glm(formula = Status ~ ., family = binomial(), data = final)  
##   
## Coefficients: (4 not defined because of singularities)  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.623197 0.659613 -0.945 0.344765   
## Age 0.030555 0.007344 4.161 3.17e-05 \*\*\*  
## RaceOther -1.021996 0.331075 -3.087 0.002023 \*\*   
## RaceWhite -0.590952 0.233724 -2.528 0.011458 \*   
## Marital.StatusMarried -0.308122 0.198315 -1.554 0.120256   
## Marital.StatusSeparated 0.667096 0.571020 1.168 0.242705   
## Marital.StatusSingle -0.218432 0.244742 -0.892 0.372127   
## Marital.StatusWidowed -0.198336 0.308707 -0.642 0.520565   
## T.StageT2 0.888868 0.269167 3.302 0.000959 \*\*\*  
## T.StageT3 1.486164 0.448708 3.312 0.000926 \*\*\*  
## T.StageT4 1.368733 0.652902 2.096 0.036048 \*   
## N.StageN2 0.867566 0.341680 2.539 0.011113 \*   
## N.StageN3 0.815172 0.413751 1.970 0.048816 \*   
## X6th.StageIIB -0.229432 0.304720 -0.753 0.451493   
## X6th.StageIIIA -0.321014 0.399554 -0.803 0.421726   
## X6th.StageIIIB 0.134539 0.760334 0.177 0.859550   
## X6th.StageIIIC NA NA NA NA   
## differentiatePoorly differentiated 0.269454 0.144471 1.865 0.062167 .   
## differentiateUndifferentiated 1.894486 1.128711 1.678 0.093259 .   
## differentiateWell differentiated -0.748134 0.222072 -3.369 0.000755 \*\*\*  
## Grade1 NA NA NA NA   
## Grade2 NA NA NA NA   
## Grade3 NA NA NA NA   
## A.StageRegional 0.501456 0.366124 1.370 0.170801   
## Tumor.Size -0.010394 0.005433 -1.913 0.055735 .   
## Estrogen.StatusPositive -0.824936 0.278089 -2.966 0.003013 \*\*   
## Progesterone.StatusPositive -0.583886 0.183948 -3.174 0.001503 \*\*   
## Regional.Node.Examined -0.020523 0.009010 -2.278 0.022736 \*   
## Reginol.Node.Positive 0.039046 0.021657 1.803 0.071403 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1707.9 on 1231 degrees of freedom  
## Residual deviance: 1442.0 on 1207 degrees of freedom  
## AIC: 1492  
##   
## Number of Fisher Scoring iterations: 5

predictions <- predict(logit\_model, type = "response")  
  
final$predicted\_classes <- ifelse(predictions > 0.5, 1, 0)  
  
confusionMatrix(as.factor(final$predicted\_classes), as.factor(final$Status))

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 451 207  
## 1 165 409  
##   
## Accuracy : 0.6981   
## 95% CI : (0.6716, 0.7236)  
## No Information Rate : 0.5   
## P-Value [Acc > NIR] : < 2e-16   
##   
## Kappa : 0.3961   
##   
## Mcnemar's Test P-Value : 0.03352   
##   
## Sensitivity : 0.7321   
## Specificity : 0.6640   
## Pos Pred Value : 0.6854   
## Neg Pred Value : 0.7125   
## Prevalence : 0.5000   
## Detection Rate : 0.3661   
## Detection Prevalence : 0.5341   
## Balanced Accuracy : 0.6981   
##   
## 'Positive' Class : 0   
##

## Assumptions

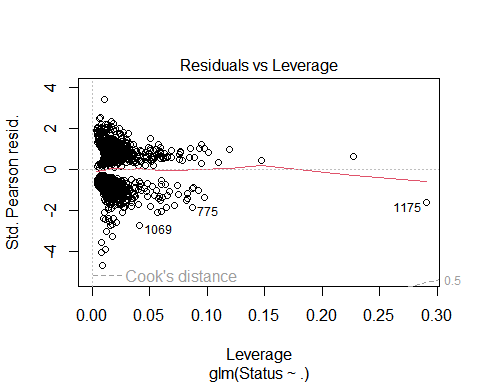
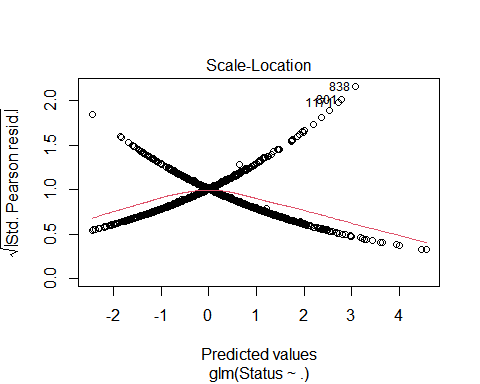
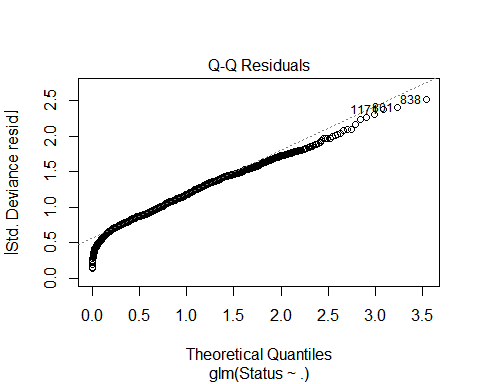
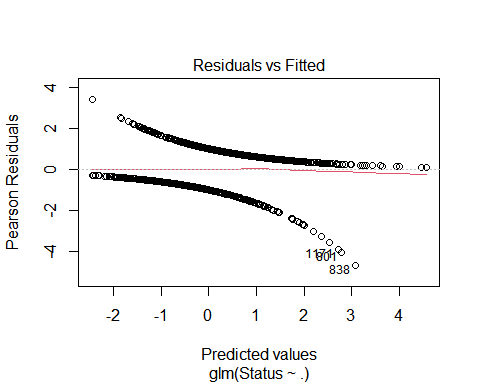
#vif(logit\_model)  
# there are aliased coefficients in the model > means some vars have perfect multicoll.   
#cor(sampledf)  
#come back to this

## Full model diagnostics

#full model LL   
logLik(logit\_model)

## 'log Lik.' -721.0144 (df=25)

#Diagnostic plots  
plot(logit\_model)

 ## Variable selection

#AIC vs. BIC var selection criteria  
#Akaike Info Criterion  
final\_aic<-bind\_rows(dead, alive) |>  
 dplyr::select(-Survival.Months)  
  
stepAIC(logit\_model, trace=0, k = 2)

##   
## Call: glm(formula = Status ~ Age + Race + T.Stage + N.Stage + differentiate +   
## Tumor.Size + Estrogen.Status + Progesterone.Status + Regional.Node.Examined +   
## Reginol.Node.Positive, family = binomial(), data = final)  
##   
## Coefficients:  
## (Intercept) Age   
## -0.38878 0.03027   
## RaceOther RaceWhite   
## -1.08741 -0.63335   
## T.StageT2 T.StageT3   
## 0.73505 1.29120   
## T.StageT4 N.StageN2   
## 1.33645 0.69474   
## N.StageN3 differentiatePoorly differentiated   
## 0.85333 0.28302   
## differentiateUndifferentiated differentiateWell differentiated   
## 1.90761 -0.73595   
## Tumor.Size Estrogen.StatusPositive   
## -0.01021 -0.78051   
## Progesterone.StatusPositive Regional.Node.Examined   
## -0.59710 -0.02131   
## Reginol.Node.Positive   
## 0.04327   
##   
## Degrees of Freedom: 1231 Total (i.e. Null); 1215 Residual  
## Null Deviance: 1708   
## Residual Deviance: 1450 AIC: 1484

#model determined by AIC criterion: minimize AIC for better model fit  
logit\_aic <- glm(Status ~ Age + Race + Marital.Status + X6th.Stage +   
 differentiate + Estrogen.Status + Progesterone.Status + Regional.Node.Examined +   
 Reginol.Node.Positive, family = binomial(), data = final\_aic)  
#log likelihood  
logLik(logit\_aic)

## 'log Lik.' -730.5563 (df=19)

vif(logit\_aic)

## GVIF Df GVIF^(1/(2\*Df))  
## Age 1.130797 1 1.063389  
## Race 1.107848 2 1.025935  
## Marital.Status 1.179101 4 1.020808  
## X6th.Stage 3.622899 4 1.174578  
## differentiate 1.092877 3 1.014912  
## Estrogen.Status 1.403712 1 1.184784  
## Progesterone.Status 1.361144 1 1.166681  
## Regional.Node.Examined 1.370111 1 1.170517  
## Reginol.Node.Positive 3.882240 1 1.970340

#confusion matrix  
predictions\_aic <- predict(logit\_aic, type = "response")  
final\_aic$predicted\_classes <- ifelse(predictions\_aic > 0.5, 1, 0)  
confusionMatrix(as.factor(final\_aic$predicted\_classes), as.factor(final\_aic$Status))

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 435 202  
## 1 181 414  
##   
## Accuracy : 0.6891   
## 95% CI : (0.6624, 0.7149)  
## No Information Rate : 0.5   
## P-Value [Acc > NIR] : <2e-16   
##   
## Kappa : 0.3782   
##   
## Mcnemar's Test P-Value : 0.3068   
##   
## Sensitivity : 0.7062   
## Specificity : 0.6721   
## Pos Pred Value : 0.6829   
## Neg Pred Value : 0.6958   
## Prevalence : 0.5000   
## Detection Rate : 0.3531   
## Detection Prevalence : 0.5170   
## Balanced Accuracy : 0.6891   
##   
## 'Positive' Class : 0   
##

#Bayesian Info Criterion  
final\_bic<-bind\_rows(dead, alive) |>  
 dplyr::select(-Survival.Months)  
  
stepAIC(logit\_model, trace=0, k = log(nrow(sampledf)))

##   
## Call: glm(formula = Status ~ Age + N.Stage + differentiate + Progesterone.Status,   
## family = binomial(), data = final)  
##   
## Coefficients:  
## (Intercept) Age   
## -1.32377 0.02744   
## N.StageN2 N.StageN3   
## 0.86704 1.38420   
## differentiatePoorly differentiated differentiateUndifferentiated   
## 0.39357 2.16872   
## differentiateWell differentiated Progesterone.StatusPositive   
## -0.71833 -0.88912   
##   
## Degrees of Freedom: 1231 Total (i.e. Null); 1224 Residual  
## Null Deviance: 1708   
## Residual Deviance: 1501 AIC: 1517

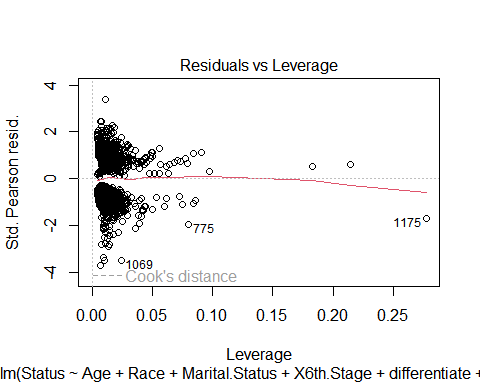
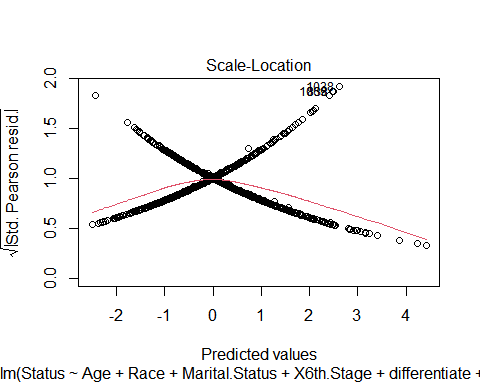
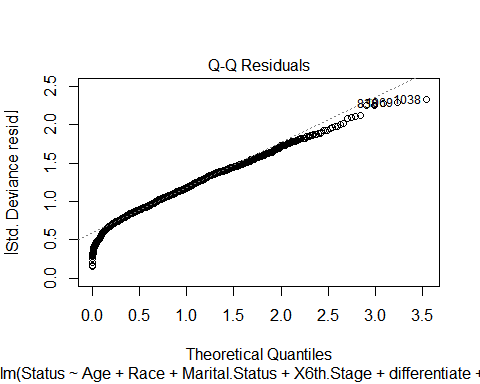
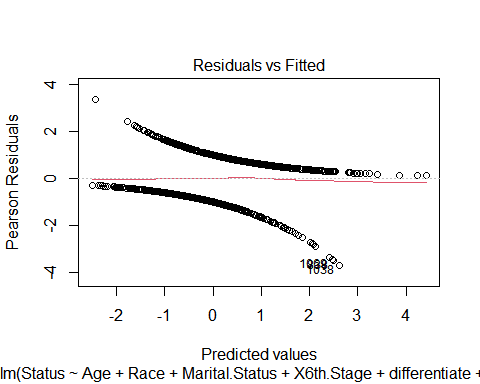
#model determined by BIC criterion: minimize BIC for model fit  
logit\_bic <- glm(Status ~ differentiate + Progesterone.Status +   
 Regional.Node.Examined + Reginol.Node.Positive, family = binomial(),   
 data = final\_bic)  
#log likelihood  
logLik(logit\_bic)

## 'log Lik.' -763.6348 (df=7)

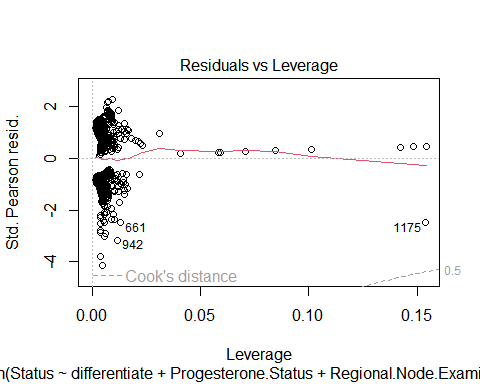
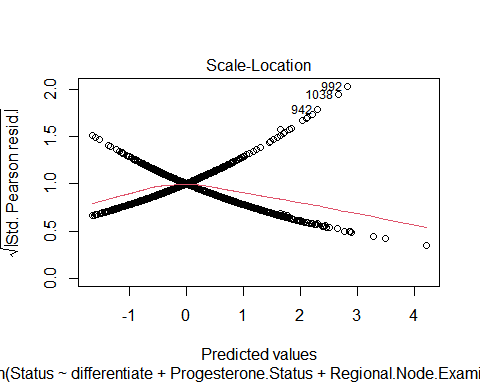
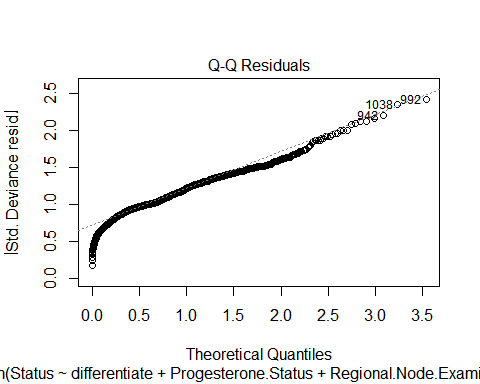
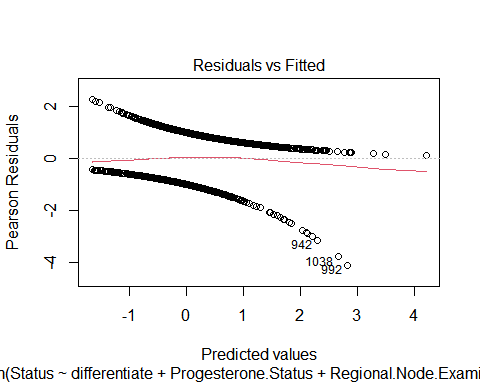
#confusion matrix  
predictions\_bic <- predict(logit\_bic, type = "response")  
final\_bic$predicted\_classes <- ifelse(predictions\_bic > 0.5, 1, 0)  
confusionMatrix(as.factor(final\_bic$predicted\_classes), as.factor(final\_bic$Status))

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 461 260  
## 1 155 356  
##   
## Accuracy : 0.6631   
## 95% CI : (0.636, 0.6895)  
## No Information Rate : 0.5   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.3263   
##   
## Mcnemar's Test P-Value : 3.305e-07   
##   
## Sensitivity : 0.7484   
## Specificity : 0.5779   
## Pos Pred Value : 0.6394   
## Neg Pred Value : 0.6967   
## Prevalence : 0.5000   
## Detection Rate : 0.3742   
## Detection Prevalence : 0.5852   
## Balanced Accuracy : 0.6631   
##   
## 'Positive' Class : 0   
##

#make diagnostic plots to address outlying observations and assumptions  
#Diagnostic plot for AIC model  
plot(logit\_aic)



#Diagnostic plot for BIC model  
plot(logit\_bic)



#refit models using sensitivity-specificity plots and then check performance and sensitivity/specificity

we can then pick and interpret the best final model

## Use the balanced model for regression

# Load required libraries  
library(tidyverse)  
library(survival)  
library(caret)  
library(car) # For VIF checks  
library(MASS) # For stepwise variable selection  
library(glmnet)

## Warning: 程序包'glmnet'是用R版本4.4.2 来建造的

## 载入需要的程序包：Matrix

##   
## 载入程序包：'Matrix'

## The following objects are masked from 'package:tidyr':  
##   
## expand, pack, unpack

## Loaded glmnet 4.1-8

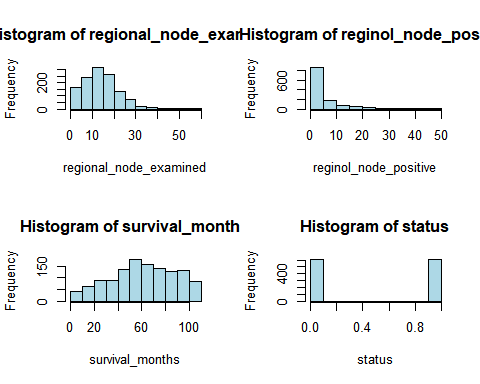
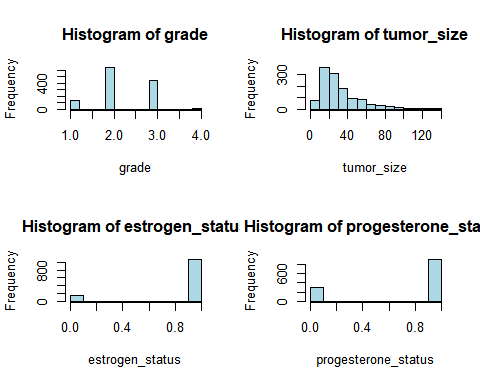
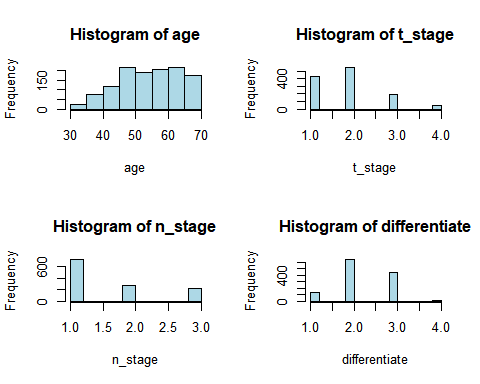
# Load the balanced dataset  
balance\_data <- read.csv("project\_2\_numdata.csv")  
  
# View dataset structure (optional)  
str(balance\_data)

## 'data.frame': 1232 obs. of 17 variables:  
## $ age : int 40 68 57 42 67 31 64 53 63 45 ...  
## $ race : chr "White" "White" "White" "White" ...  
## $ marital\_status : chr "Married" "Widowed" "Single" "Married" ...  
## $ t\_stage : int 2 1 3 1 1 3 2 3 2 2 ...  
## $ n\_stage : int 1 1 3 3 1 3 3 1 2 3 ...  
## $ x6th\_stage : chr "IIB" "IIA" "IIIC" "IIIC" ...  
## $ differentiate : int 2 2 2 2 2 3 3 3 1 3 ...  
## $ grade : int 2 2 2 2 2 3 3 3 1 3 ...  
## $ a\_stage : chr "Regional" "Regional" "Regional" "Regional" ...  
## $ tumor\_size : int 30 13 70 9 13 70 34 60 26 25 ...  
## $ estrogen\_status : int 1 1 1 0 1 1 1 1 1 1 ...  
## $ progesterone\_status : int 1 1 1 0 1 1 0 1 1 1 ...  
## $ regional\_node\_examined: int 9 9 12 15 16 23 24 8 5 21 ...  
## $ reginol\_node\_positive : int 1 1 12 2 3 23 20 1 4 16 ...  
## $ survival\_months : int 14 64 42 39 60 44 12 91 87 40 ...  
## $ status : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ a\_stage\_regional : int 1 1 1 1 1 1 1 1 1 1 ...

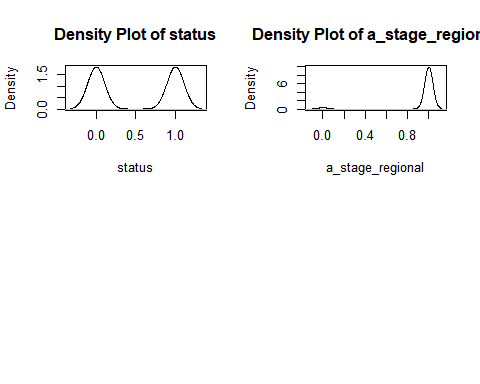
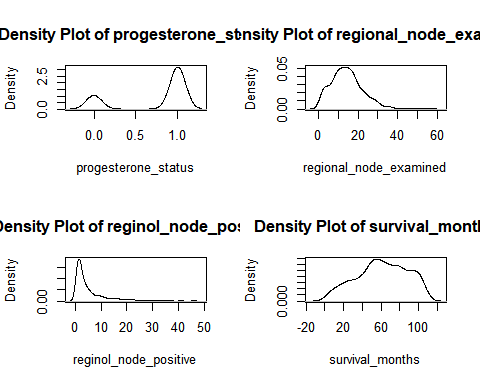
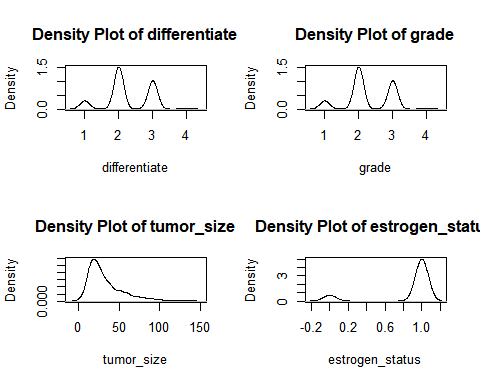
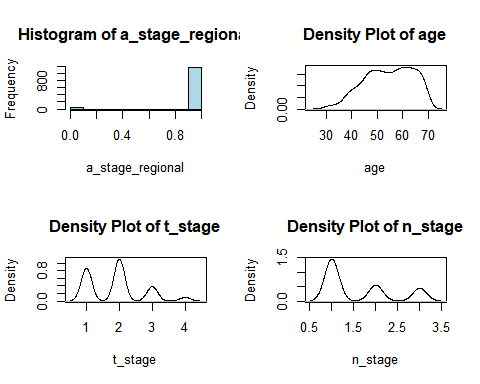
view(balance\_data)  
  
# Summary statistics for numerical variables  
summary(balance\_data)

## age race marital\_status t\_stage   
## Min. :30.0 Length:1232 Length:1232 Min. :1.000   
## 1st Qu.:47.0 Class :character Class :character 1st Qu.:1.000   
## Median :55.0 Mode :character Mode :character Median :2.000   
## Mean :54.5 Mean :1.881   
## 3rd Qu.:62.0 3rd Qu.:2.000   
## Max. :69.0 Max. :4.000   
## n\_stage x6th\_stage differentiate grade   
## Min. :1.000 Length:1232 Min. :1.000 Min. :1.000   
## 1st Qu.:1.000 Class :character 1st Qu.:2.000 1st Qu.:2.000   
## Median :1.000 Mode :character Median :2.000 Median :2.000   
## Mean :1.587 Mean :2.269 Mean :2.269   
## 3rd Qu.:2.000 3rd Qu.:3.000 3rd Qu.:3.000   
## Max. :3.000 Max. :4.000 Max. :4.000   
## a\_stage tumor\_size estrogen\_status progesterone\_status  
## Length:1232 Min. : 1.0 Min. :0.0000 Min. :0.0000   
## Class :character 1st Qu.: 17.0 1st Qu.:1.0000 1st Qu.:1.0000   
## Mode :character Median : 25.0 Median :1.0000 Median :1.0000   
## Mean : 32.8 Mean :0.8815 Mean :0.7549   
## 3rd Qu.: 40.0 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :140.0 Max. :1.0000 Max. :1.0000   
## regional\_node\_examined reginol\_node\_positive survival\_months status   
## Min. : 1.00 Min. : 1.000 Min. : 2.00 Min. :0.0   
## 1st Qu.: 9.00 1st Qu.: 1.000 1st Qu.: 43.00 1st Qu.:0.0   
## Median :14.00 Median : 3.000 Median : 62.00 Median :0.5   
## Mean :14.52 Mean : 5.281 Mean : 61.24 Mean :0.5   
## 3rd Qu.:19.00 3rd Qu.: 7.000 3rd Qu.: 82.00 3rd Qu.:1.0   
## Max. :57.00 Max. :46.000 Max. :107.00 Max. :1.0   
## a\_stage\_regional  
## Min. :0.0000   
## 1st Qu.:1.0000   
## Median :1.0000   
## Mean :0.9667   
## 3rd Qu.:1.0000   
## Max. :1.0000

# Visualize distributions of numerical variables  
numerical\_vars <- balance\_data[, sapply(balance\_data, is.numeric)]  
  
# Histograms  
par(mfrow = c(2, 2)) # Arrange multiple plots  
for (col in colnames(numerical\_vars)) {  
 hist(numerical\_vars[[col]], main = paste("Histogram of", col), xlab = col, col = "lightblue")  
}



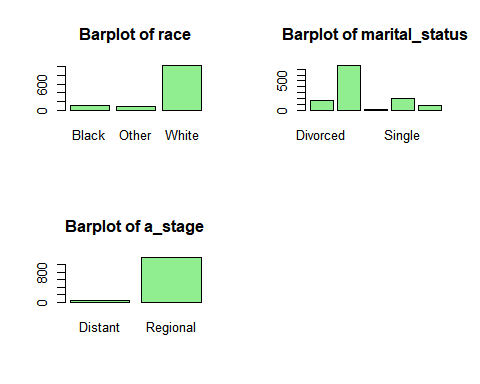
# Density plots for more detailed distributions  
for (col in colnames(numerical\_vars)) {  
 plot(density(numerical\_vars[[col]], na.rm = TRUE), main = paste("Density Plot of", col), xlab = col)  
}



balance\_data$race <- as.factor(balance\_data$race)  
balance\_data$marital\_status <- as.factor(balance\_data$marital\_status)  
balance\_data$a\_stage <- as.factor(balance\_data$a\_stage)  
  
# Frequency tables  
categorical\_vars <- balance\_data[, sapply(balance\_data, is.factor)]  
  
for (col in colnames(categorical\_vars)) {  
 cat("Frequency of", col, ":\n")  
 print(table(categorical\_vars[[col]]))  
}

## Frequency of race :  
##   
## Black Other White   
## 118 79 1035   
## Frequency of marital\_status :  
##   
## Divorced Married Separated Single Widowed   
## 158 765 21 201 87   
## Frequency of a\_stage :  
##   
## Distant Regional   
## 41 1191

# Bar plots  
par(mfrow = c(2, 2))  
for (col in colnames(categorical\_vars)) {  
 barplot(table(categorical\_vars[[col]]), main = paste("Barplot of", col), col = "lightgreen")  
}



library(GGally)

## Warning: 程序包'GGally'是用R版本4.4.2 来建造的

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

# Correlation matrix  
cor\_matrix <- cor(balance\_data[, sapply(balance\_data, is.numeric)], use = "complete.obs")  
print(cor\_matrix)

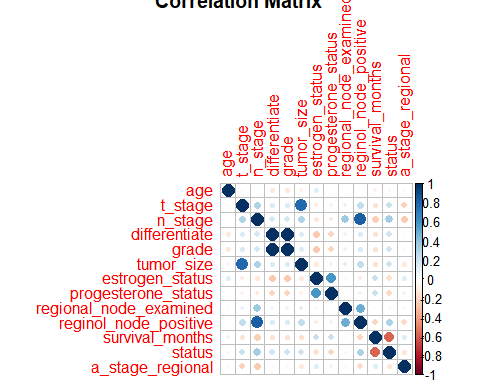
## age t\_stage n\_stage differentiate  
## age 1.00000000 -0.05011340 -0.01706295 -0.12321448  
## t\_stage -0.05011340 1.00000000 0.30612098 0.15076838  
## n\_stage -0.01706295 0.30612098 1.00000000 0.17991601  
## differentiate -0.12321448 0.15076838 0.17991601 1.00000000  
## grade -0.12321448 0.15076838 0.17991601 1.00000000  
## tumor\_size -0.07953275 0.78499755 0.30277440 0.14403058  
## estrogen\_status 0.12111658 -0.10672267 -0.12700846 -0.26247316  
## progesterone\_status 0.01233580 -0.07197461 -0.11108594 -0.21201525  
## regional\_node\_examined -0.03141638 0.08844292 0.36808270 0.07651544  
## reginol\_node\_positive 0.01628318 0.24426697 0.82874594 0.15163577  
## survival\_months -0.08599693 -0.15971602 -0.24737291 -0.15052637  
## status 0.06976636 0.22525091 0.34556514 0.19151006  
## a\_stage\_regional 0.04777171 -0.23467556 -0.25575034 -0.04874199  
## grade tumor\_size estrogen\_status  
## age -0.12321448 -0.07953275 0.1211166  
## t\_stage 0.15076838 0.78499755 -0.1067227  
## n\_stage 0.17991601 0.30277440 -0.1270085  
## differentiate 1.00000000 0.14403058 -0.2624732  
## grade 1.00000000 0.14403058 -0.2624732  
## tumor\_size 0.14403058 1.00000000 -0.1230207  
## estrogen\_status -0.26247316 -0.12302067 1.0000000  
## progesterone\_status -0.21201525 -0.08438705 0.5850460  
## regional\_node\_examined 0.07651544 0.11473470 -0.0414043  
## reginol\_node\_positive 0.15163577 0.24878553 -0.1008736  
## survival\_months -0.15052637 -0.16151178 0.2272740  
## status 0.19151006 0.19535827 -0.1757951  
## a\_stage\_regional -0.04874199 -0.12796649 0.1279901  
## progesterone\_status regional\_node\_examined  
## age 0.01233580 -0.03141638  
## t\_stage -0.07197461 0.08844292  
## n\_stage -0.11108594 0.36808270  
## differentiate -0.21201525 0.07651544  
## grade -0.21201525 0.07651544  
## tumor\_size -0.08438705 0.11473470  
## estrogen\_status 0.58504596 -0.04140430  
## progesterone\_status 1.00000000 -0.01390137  
## regional\_node\_examined -0.01390137 1.00000000  
## reginol\_node\_positive -0.07105526 0.49395643  
## survival\_months 0.21200308 -0.03087185  
## status -0.20001424 0.06199862  
## a\_stage\_regional 0.07311096 -0.04237809  
## reginol\_node\_positive survival\_months status  
## age 0.01628318 -0.08599693 0.06976636  
## t\_stage 0.24426697 -0.15971602 0.22525091  
## n\_stage 0.82874594 -0.24737291 0.34556514  
## differentiate 0.15163577 -0.15052637 0.19151006  
## grade 0.15163577 -0.15052637 0.19151006  
## tumor\_size 0.24878553 -0.16151178 0.19535827  
## estrogen\_status -0.10087358 0.22727403 -0.17579507  
## progesterone\_status -0.07105526 0.21200308 -0.20001424  
## regional\_node\_examined 0.49395643 -0.03087185 0.06199862  
## reginol\_node\_positive 1.00000000 -0.21263747 0.31134369  
## survival\_months -0.21263747 1.00000000 -0.58302039  
## status 0.31134369 -0.58302039 1.00000000  
## a\_stage\_regional -0.19505294 0.13738966 -0.13123516  
## a\_stage\_regional  
## age 0.04777171  
## t\_stage -0.23467556  
## n\_stage -0.25575034  
## differentiate -0.04874199  
## grade -0.04874199  
## tumor\_size -0.12796649  
## estrogen\_status 0.12799010  
## progesterone\_status 0.07311096  
## regional\_node\_examined -0.04237809  
## reginol\_node\_positive -0.19505294  
## survival\_months 0.13738966  
## status -0.13123516  
## a\_stage\_regional 1.00000000

# Visualize correlations  
library(corrplot)

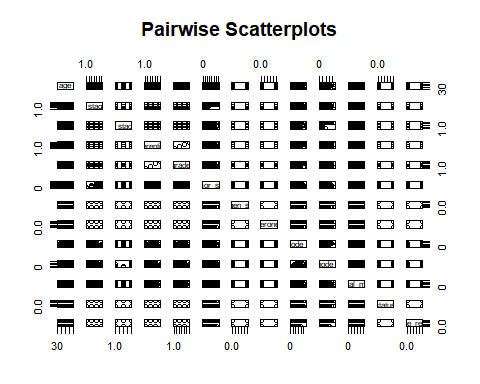
## Warning: 程序包'corrplot'是用R版本4.4.2 来建造的

## corrplot 0.95 loaded

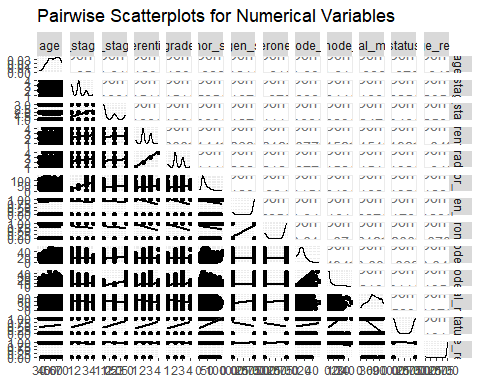
corrplot(cor\_matrix, method = "circle", main = "Correlation Matrix")



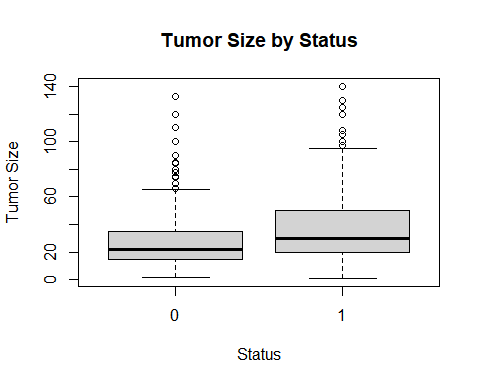
# Pairwise scatterplots for key numerical variables  
pairs(numerical\_vars, main = "Pairwise Scatterplots")



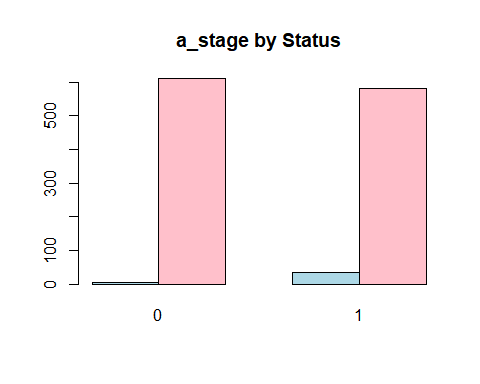
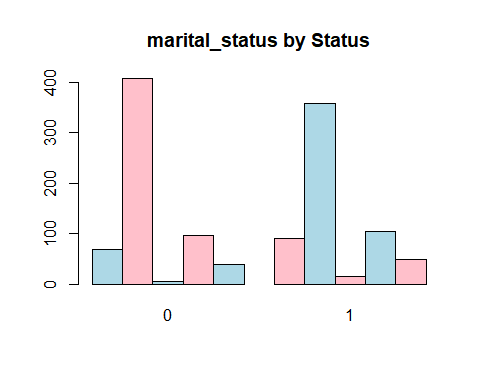
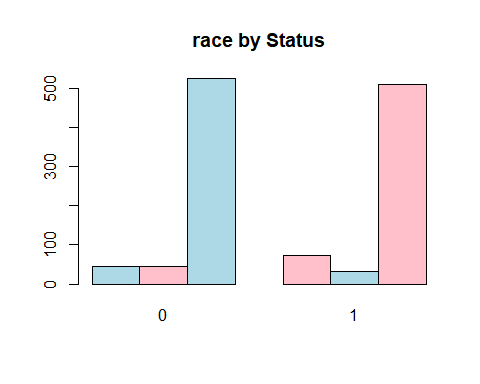
ggpairs(numerical\_vars,   
 title = "Pairwise Scatterplots for Numerical Variables",  
 upper = list(continuous = wrap("cor", size = 4)), # Add correlation in upper triangle  
 lower = list(continuous = "smooth"), # Scatterplot with smooth line  
 diag = list(continuous = "densityDiag")) # Density plot on diagonal



# Boxplot to examine relationship between a numerical variable and Status  
boxplot(balance\_data$tumor\_size ~ balance\_data$status, main = "Tumor Size by Status", xlab = "Status", ylab = "Tumor Size")



# Barplot for categorical variables vs. Status  
for (col in colnames(categorical\_vars)) {  
 table\_by\_status <- table(balance\_data[[col]], balance\_data$status)  
 barplot(table\_by\_status, beside = TRUE, main = paste(col, "by Status"), col = c("lightblue", "pink"))  
}



# Logistic Regression Full Model  
logit\_model\_balance <- glm(status ~ ., data = balance\_data, family = binomial())  
  
# Model Summary  
summary(logit\_model\_balance)

##   
## Call:  
## glm(formula = status ~ ., family = binomial(), data = balance\_data)  
##   
## Coefficients: (2 not defined because of singularities)  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.211557 1.057531 1.146 0.25194   
## age 0.014887 0.008731 1.705 0.08819 .   
## raceOther -0.434926 0.415309 -1.047 0.29499   
## raceWhite -0.425415 0.286581 -1.484 0.13769   
## marital\_statusMarried -0.449173 0.239708 -1.874 0.06095 .   
## marital\_statusSeparated 0.216378 0.712169 0.304 0.76126   
## marital\_statusSingle -0.359626 0.299139 -1.202 0.22928   
## marital\_statusWidowed -0.200391 0.367155 -0.546 0.58521   
## t\_stage 0.253031 0.230527 1.098 0.27237   
## n\_stage 1.009773 0.374068 2.699 0.00695 \*\*   
## x6th\_stageIIB 0.331425 0.269520 1.230 0.21881   
## x6th\_stageIIIA -0.249454 0.455305 -0.548 0.58377   
## x6th\_stageIIIB 0.660541 0.900261 0.734 0.46312   
## x6th\_stageIIIC -1.312856 0.911535 -1.440 0.14979   
## differentiate 0.379493 0.125072 3.034 0.00241 \*\*   
## grade NA NA NA NA   
## a\_stageRegional 0.003970 0.568085 0.007 0.99442   
## tumor\_size 0.001213 0.006669 0.182 0.85563   
## estrogen\_status -0.044378 0.335420 -0.132 0.89474   
## progesterone\_status -0.491775 0.223295 -2.202 0.02764 \*   
## regional\_node\_examined -0.025331 0.010863 -2.332 0.01971 \*   
## reginol\_node\_positive 0.082204 0.028517 2.883 0.00394 \*\*   
## survival\_months -0.058808 0.003867 -15.209 < 2e-16 \*\*\*  
## a\_stage\_regional NA NA NA NA   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1707.9 on 1231 degrees of freedom  
## Residual deviance: 1076.3 on 1210 degrees of freedom  
## AIC: 1120.3  
##   
## Number of Fisher Scoring iterations: 5

# Predictions  
predictions <- predict(logit\_model\_balance, type = "response")  
  
# Classify Predictions  
balance\_data$predicted\_classes <- ifelse(predictions > 0.5, 1, 0)  
  
# Ensure consistent factor levels  
balance\_data$predicted\_classes <- factor(balance\_data$predicted\_classes, levels = c(0, 1))  
balance\_data$status <- factor(balance\_data$status, levels = c(0, 1))  
  
# Confusion Matrix  
confusionMatrix(balance\_data$predicted\_classes, balance\_data$status)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 498 138  
## 1 118 478  
##   
## Accuracy : 0.7922   
## 95% CI : (0.7685, 0.8146)  
## No Information Rate : 0.5   
## P-Value [Acc > NIR] : <2e-16   
##   
## Kappa : 0.5844   
##   
## Mcnemar's Test P-Value : 0.235   
##   
## Sensitivity : 0.8084   
## Specificity : 0.7760   
## Pos Pred Value : 0.7830   
## Neg Pred Value : 0.8020   
## Prevalence : 0.5000   
## Detection Rate : 0.4042   
## Detection Prevalence : 0.5162   
## Balanced Accuracy : 0.7922   
##   
## 'Positive' Class : 0   
##

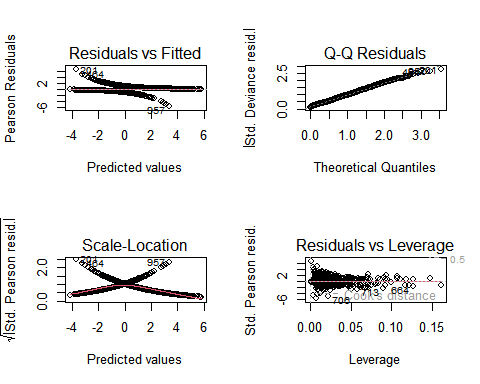
# Variance Inflation Factor (VIF) to detect multicollinearity  
library(car)  
#vif(logit\_model\_balance)  
  
# Correlation Matrix  
cor\_matrix <- cor(balance\_data[, sapply(balance\_data, is.numeric)], use = "complete.obs")  
print(cor\_matrix)

## age t\_stage n\_stage differentiate  
## age 1.00000000 -0.05011340 -0.01706295 -0.12321448  
## t\_stage -0.05011340 1.00000000 0.30612098 0.15076838  
## n\_stage -0.01706295 0.30612098 1.00000000 0.17991601  
## differentiate -0.12321448 0.15076838 0.17991601 1.00000000  
## grade -0.12321448 0.15076838 0.17991601 1.00000000  
## tumor\_size -0.07953275 0.78499755 0.30277440 0.14403058  
## estrogen\_status 0.12111658 -0.10672267 -0.12700846 -0.26247316  
## progesterone\_status 0.01233580 -0.07197461 -0.11108594 -0.21201525  
## regional\_node\_examined -0.03141638 0.08844292 0.36808270 0.07651544  
## reginol\_node\_positive 0.01628318 0.24426697 0.82874594 0.15163577  
## survival\_months -0.08599693 -0.15971602 -0.24737291 -0.15052637  
## a\_stage\_regional 0.04777171 -0.23467556 -0.25575034 -0.04874199  
## grade tumor\_size estrogen\_status  
## age -0.12321448 -0.07953275 0.1211166  
## t\_stage 0.15076838 0.78499755 -0.1067227  
## n\_stage 0.17991601 0.30277440 -0.1270085  
## differentiate 1.00000000 0.14403058 -0.2624732  
## grade 1.00000000 0.14403058 -0.2624732  
## tumor\_size 0.14403058 1.00000000 -0.1230207  
## estrogen\_status -0.26247316 -0.12302067 1.0000000  
## progesterone\_status -0.21201525 -0.08438705 0.5850460  
## regional\_node\_examined 0.07651544 0.11473470 -0.0414043  
## reginol\_node\_positive 0.15163577 0.24878553 -0.1008736  
## survival\_months -0.15052637 -0.16151178 0.2272740  
## a\_stage\_regional -0.04874199 -0.12796649 0.1279901  
## progesterone\_status regional\_node\_examined  
## age 0.01233580 -0.03141638  
## t\_stage -0.07197461 0.08844292  
## n\_stage -0.11108594 0.36808270  
## differentiate -0.21201525 0.07651544  
## grade -0.21201525 0.07651544  
## tumor\_size -0.08438705 0.11473470  
## estrogen\_status 0.58504596 -0.04140430  
## progesterone\_status 1.00000000 -0.01390137  
## regional\_node\_examined -0.01390137 1.00000000  
## reginol\_node\_positive -0.07105526 0.49395643  
## survival\_months 0.21200308 -0.03087185  
## a\_stage\_regional 0.07311096 -0.04237809  
## reginol\_node\_positive survival\_months a\_stage\_regional  
## age 0.01628318 -0.08599693 0.04777171  
## t\_stage 0.24426697 -0.15971602 -0.23467556  
## n\_stage 0.82874594 -0.24737291 -0.25575034  
## differentiate 0.15163577 -0.15052637 -0.04874199  
## grade 0.15163577 -0.15052637 -0.04874199  
## tumor\_size 0.24878553 -0.16151178 -0.12796649  
## estrogen\_status -0.10087358 0.22727403 0.12799010  
## progesterone\_status -0.07105526 0.21200308 0.07311096  
## regional\_node\_examined 0.49395643 -0.03087185 -0.04237809  
## reginol\_node\_positive 1.00000000 -0.21263747 -0.19505294  
## survival\_months -0.21263747 1.00000000 0.13738966  
## a\_stage\_regional -0.19505294 0.13738966 1.00000000

# Log-Likelihood of the Full Model  
logLik(logit\_model\_balance)

## 'log Lik.' -538.1702 (df=22)

# Diagnostic Plots  
par(mfrow = c(2, 2))  
plot(logit\_model\_balance)



par(mfrow = c(1, 1)) # Reset graphical parameters

# Stepwise Selection using AIC

logit\_aic <- stepAIC(logit\_model, trace = 0, k = 2)

# Final AIC Model Summary

summary(logit\_aic)

# Log-Likelihood and VIF

logLik(logit\_aic) library(car) vif(logit\_aic)

# Predictions

predictions\_aic <- predict(logit\_aic, type = “response”)

# Classify Predictions

data$predicted\_classes\_aic <- ifelse(predictions\_aic > 0.5, 1, 0)

# Ensure factors with consistent levels

datapredicted\_classes\_aic, levels = c(0, 1)) datastatus, levels = c(0, 1))

# Confusion Matrix

library(caret) confusionMatrix(datastatus)

# Stepwise Selection using BIC  
logit\_bic <- stepAIC(logit\_model, trace = 0, k = log(nrow(data)))  
  
# Final BIC Model Summary  
summary(logit\_bic)  
  
# Log-Likelihood  
logLik(logit\_bic)  
  
# Predictions  
predictions\_bic <- predict(logit\_bic, type = "response")  
  
# Classify Predictions  
data$predicted\_classes\_bic <- ifelse(predictions\_bic > 0.5, 1, 0)  
  
# Ensure factors with consistent levels  
data$predicted\_classes\_bic <- factor(data$predicted\_classes\_bic, levels = c(0, 1))  
data$Status <- factor(data$Status, levels = c(0, 1))  
  
# Confusion Matrix  
library(caret)  
confusionMatrix(data$predicted\_classes\_bic, data$Status)

# Log-Likelihood for AIC and BIC Models

logLik(logit\_aic) logLik(logit\_bic)

# Predictions and Classification for AIC Model

predictions\_aic <- predict(logit\_aic, type = “response”) balance\_data$predicted\_classes\_aic <- ifelse(predictions\_aic > 0.5, 1, 0)

# Predictions and Classification for BIC Model

predictions\_bic <- predict(logit\_bic, type = “response”) balance\_data$predicted\_classes\_bic <- ifelse(predictions\_bic > 0.5, 1, 0)

# Ensure consistent factor levels for both AIC and BIC

balance\_datapredicted\_classes\_aic, levels = c(0, 1)) balance\_datapredicted\_classes\_bic, levels = c(0, 1)) balance\_datastatus, levels = c(0, 1))

# Confusion Matrices

library(caret) confusionMatrix(balance\_datastatus) confusionMatrix(balance\_datastatus)

# AUC-ROC for AIC and BIC Models

library(pROC) roc\_curve\_aic <- roc(as.numeric(as.character(balance\_dataStatus)), predictions\_bic)

# Plot ROC Curves

plot(roc\_curve\_aic, col = “blue”, main = “ROC Curve for AIC and BIC Models”, lwd = 2) lines(roc\_curve\_bic, col = “red”, lwd = 2) legend(“bottomright”, legend = c(“AIC”, “BIC”), col = c(“blue”, “red”), lwd = 2)

# AUC for AIC and BIC Models

auc(roc\_curve\_aic) auc(roc\_curve\_bic)

## Interaction models!!!  
  
  
``` r  
library(fastDummies)

## Warning: 程序包'fastDummies'是用R版本4.4.2 来建造的

# Create dummy variables for categorical variables  
balance\_data\_dummies <- dummy\_cols(  
 balance\_data,  
 select\_columns = c("race", "marital\_status", "x6th\_stage", "a\_stage"),  
 remove\_first\_dummy = TRUE, # Avoid multicollinearity by removing the first dummy  
 remove\_selected\_columns = TRUE # Remove original columns  
)  
  
# View the updated dataset  
head(balance\_data\_dummies)

## age t\_stage n\_stage differentiate grade tumor\_size estrogen\_status  
## 1 40 2 1 2 2 30 1  
## 2 68 1 1 2 2 13 1  
## 3 57 3 3 2 2 70 1  
## 4 42 1 3 2 2 9 0  
## 5 67 1 1 2 2 13 1  
## 6 31 3 3 3 3 70 1  
## progesterone\_status regional\_node\_examined reginol\_node\_positive  
## 1 1 9 1  
## 2 1 9 1  
## 3 1 12 12  
## 4 0 15 2  
## 5 1 16 3  
## 6 1 23 23  
## survival\_months status a\_stage\_regional predicted\_classes race\_Other  
## 1 14 1 1 1 0  
## 2 64 1 1 0 0  
## 3 42 1 1 1 0  
## 4 39 1 1 1 0  
## 5 60 1 1 0 0  
## 6 44 1 1 1 0  
## race\_White marital\_status\_Married marital\_status\_Separated  
## 1 1 1 0  
## 2 1 0 0  
## 3 1 0 0  
## 4 1 1 0  
## 5 1 0 0  
## 6 1 1 0  
## marital\_status\_Single marital\_status\_Widowed x6th\_stage\_IIB x6th\_stage\_IIIA  
## 1 0 0 1 0  
## 2 0 1 0 0  
## 3 1 0 0 0  
## 4 0 0 0 0  
## 5 0 0 0 0  
## 6 0 0 0 0  
## x6th\_stage\_IIIB x6th\_stage\_IIIC a\_stage\_Regional  
## 1 0 0 1  
## 2 0 0 1  
## 3 0 1 1  
## 4 0 1 1  
## 5 0 0 1  
## 6 0 1 1

library(glmnet)  
  
# Load required libraries  
library(MASS)  
  
# Generate all two-way interactions  
formula\_with\_interactions <- status ~   
 (age + grade + tumor\_size + estrogen\_status + progesterone\_status +   
 regional\_node\_examined + reginol\_node\_positive)^2+ +   
 marital\_status\_Single + marital\_status\_Widowed + marital\_status\_Separated +   
 x6th\_stage\_IIB + x6th\_stage\_IIIA + x6th\_stage\_IIIB + x6th\_stage\_IIIC +   
 a\_stage\_regional  
# Fit the logistic regression model with the updated formula  
interaction\_model <- glm(formula\_with\_interactions, data = balance\_data\_dummies, family = binomial())  
  
# View the summary of the model  
summary(interaction\_model)

##   
## Call:  
## glm(formula = formula\_with\_interactions, family = binomial(),   
## data = balance\_data\_dummies)  
##   
## Coefficients:  
## Estimate Std. Error z value  
## (Intercept) -5.0315671 2.7764165 -1.812  
## age 0.0612314 0.0422038 1.451  
## grade 0.8686971 0.8222391 1.057  
## tumor\_size 0.0480101 0.0301414 1.593  
## estrogen\_status 0.7586018 2.3413419 0.324  
## progesterone\_status -1.6544489 1.8996219 -0.871  
## regional\_node\_examined 0.0313134 0.0822064 0.381  
## reginol\_node\_positive 0.2185913 0.1343681 1.627  
## marital\_status\_Single 0.2066461 0.1787612 1.156  
## marital\_status\_Widowed 0.2325247 0.2630843 0.884  
## marital\_status\_Separated 0.8567306 0.5595292 1.531  
## x6th\_stage\_IIB 0.5730647 0.1991393 2.878  
## x6th\_stage\_IIIA 0.7328995 0.2651780 2.764  
## x6th\_stage\_IIIB 1.6098447 0.5854185 2.750  
## x6th\_stage\_IIIC 1.0192877 0.4531758 2.249  
## a\_stage\_regional -0.3363202 0.5100840 -0.659  
## age:grade -0.0100271 0.0119662 -0.838  
## age:tumor\_size -0.0002938 0.0003529 -0.833  
## age:estrogen\_status 0.0054306 0.0318613 0.170  
## age:progesterone\_status 0.0078672 0.0230871 0.341  
## age:regional\_node\_examined -0.0007101 0.0010501 -0.676  
## age:reginol\_node\_positive -0.0008303 0.0016078 -0.516  
## grade:tumor\_size 0.0030330 0.0055557 0.546  
## grade:estrogen\_status -0.0289540 0.4897073 -0.059  
## grade:progesterone\_status 0.1601791 0.3183625 0.503  
## grade:regional\_node\_examined -0.0033850 0.0144429 -0.234  
## grade:reginol\_node\_positive -0.0145921 0.0220955 -0.660  
## tumor\_size:estrogen\_status -0.0513603 0.0176299 -2.913  
## tumor\_size:progesterone\_status 0.0096456 0.0096120 1.003  
## tumor\_size:regional\_node\_examined -0.0001558 0.0004738 -0.329  
## tumor\_size:reginol\_node\_positive 0.0010221 0.0007258 1.408  
## estrogen\_status:progesterone\_status 0.2201728 0.8695034 0.253  
## estrogen\_status:regional\_node\_examined 0.0040641 0.0389264 0.104  
## estrogen\_status:reginol\_node\_positive 0.0327185 0.0735757 0.445  
## progesterone\_status:regional\_node\_examined -0.0062403 0.0258520 -0.241  
## progesterone\_status:reginol\_node\_positive -0.0627842 0.0551059 -1.139  
## regional\_node\_examined:reginol\_node\_positive -0.0023211 0.0013192 -1.760  
## Pr(>|z|)   
## (Intercept) 0.06995 .   
## age 0.14682   
## grade 0.29074   
## tumor\_size 0.11120   
## estrogen\_status 0.74594   
## progesterone\_status 0.38379   
## regional\_node\_examined 0.70327   
## reginol\_node\_positive 0.10378   
## marital\_status\_Single 0.24769   
## marital\_status\_Widowed 0.37678   
## marital\_status\_Separated 0.12573   
## x6th\_stage\_IIB 0.00401 \*\*  
## x6th\_stage\_IIIA 0.00571 \*\*  
## x6th\_stage\_IIIB 0.00596 \*\*  
## x6th\_stage\_IIIC 0.02450 \*   
## a\_stage\_regional 0.50968   
## age:grade 0.40206   
## age:tumor\_size 0.40513   
## age:estrogen\_status 0.86466   
## age:progesterone\_status 0.73328   
## age:regional\_node\_examined 0.49893   
## age:reginol\_node\_positive 0.60554   
## grade:tumor\_size 0.58512   
## grade:estrogen\_status 0.95285   
## grade:progesterone\_status 0.61487   
## grade:regional\_node\_examined 0.81470   
## grade:reginol\_node\_positive 0.50899   
## tumor\_size:estrogen\_status 0.00358 \*\*  
## tumor\_size:progesterone\_status 0.31562   
## tumor\_size:regional\_node\_examined 0.74231   
## tumor\_size:reginol\_node\_positive 0.15908   
## estrogen\_status:progesterone\_status 0.80010   
## estrogen\_status:regional\_node\_examined 0.91685   
## estrogen\_status:reginol\_node\_positive 0.65654   
## progesterone\_status:regional\_node\_examined 0.80926   
## progesterone\_status:reginol\_node\_positive 0.25456   
## regional\_node\_examined:reginol\_node\_positive 0.07849 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1707.9 on 1231 degrees of freedom  
## Residual deviance: 1417.4 on 1195 degrees of freedom  
## AIC: 1491.4  
##   
## Number of Fisher Scoring iterations: 6

# Perform Stepwise Selection (AIC)  
aic\_forward\_model <- stepAIC(interaction\_model, direction = "forward", trace = FALSE)  
aic\_backward\_model <- stepAIC(interaction\_model, direction = "backward", trace = FALSE)  
aic\_both\_model <- stepAIC(interaction\_model, direction = "both", trace = FALSE)  
  
# Perform Stepwise Selection (BIC)  
bic\_forward\_model <- stepAIC(interaction\_model, direction = "forward", trace = FALSE, k = log(nrow(balance\_data)))  
bic\_backward\_model <- stepAIC(interaction\_model, direction = "backward", trace = FALSE, k = log(nrow(balance\_data)))  
bic\_both\_model <- stepAIC(interaction\_model, direction = "both", trace = FALSE, k = log(nrow(balance\_data)))

# Extract Coefficients for All Models  
extract\_coefficients <- function(model, model\_name) {  
 summary <- summary(model)  
 coef <- as.data.frame(summary$coefficients)  
 coef$Term <- rownames(coef)  
 coef$Model <- model\_name  
 return(coef)  
}  
  
# Create Coefficient Tables  
aic\_forward\_coef <- extract\_coefficients(aic\_forward\_model, "AIC Forward")  
aic\_backward\_coef <- extract\_coefficients(aic\_backward\_model, "AIC Backward")  
aic\_both\_coef <- extract\_coefficients(aic\_both\_model, "AIC Both")  
bic\_forward\_coef <- extract\_coefficients(bic\_forward\_model, "BIC Forward")  
bic\_backward\_coef <- extract\_coefficients(bic\_backward\_model, "BIC Backward")  
bic\_both\_coef <- extract\_coefficients(bic\_both\_model, "BIC Both")  
  
# Merge Results for AIC Models  
aic\_combined <- merge(  
 aic\_forward\_coef[, c("Term", "Estimate", "Std. Error", "z value", "Pr(>|z|)")],  
 aic\_backward\_coef[, c("Term", "Estimate", "Std. Error", "z value", "Pr(>|z|)")],  
 by = "Term", all = TRUE, suffixes = c("\_AIC\_Forward", "\_AIC\_Backward")  
)  
aic\_combined <- merge(  
 aic\_combined,  
 aic\_both\_coef[, c("Term", "Estimate", "Std. Error", "z value", "Pr(>|z|)")],  
 by = "Term", all = TRUE  
)  
colnames(aic\_combined)[5:8] <- paste0(colnames(aic\_combined)[5:8], "\_AIC\_Both")  
  
# Merge Results for BIC Models  
bic\_combined <- merge(  
 bic\_forward\_coef[, c("Term", "Estimate", "Std. Error", "z value", "Pr(>|z|)")],  
 bic\_backward\_coef[, c("Term", "Estimate", "Std. Error", "z value", "Pr(>|z|)")],  
 by = "Term", all = TRUE, suffixes = c("\_BIC\_Forward", "\_BIC\_Backward")  
)  
bic\_combined <- merge(  
 bic\_combined,  
 bic\_both\_coef[, c("Term", "Estimate", "Std. Error", "z value", "Pr(>|z|)")],  
 by = "Term", all = TRUE  
)  
colnames(bic\_combined)[5:8] <- paste0(colnames(bic\_combined)[5:8], "\_BIC\_Both")  
  
# Combine AIC and BIC Results  
comparison\_table <- merge(aic\_combined, bic\_combined, by = "Term", all = TRUE)  
  
# Replace NA with 0 for consistency (optional)  
comparison\_table[is.na(comparison\_table)] <- 0  
library(writexl)

## Warning: 程序包'writexl'是用R版本4.4.2 来建造的

# Export the Comparison Table to Excel  
write\_xlsx(comparison\_table, "aic\_bic\_stepwise\_comparison.xlsx")  
  
# Confirm Export  
cat("The stepwise comparison table has been successfully written to 'aic\_bic\_stepwise\_comparison.xlsx'")

## The stepwise comparison table has been successfully written to 'aic\_bic\_stepwise\_comparison.xlsx'

library(pROC)

## Warning: 程序包'pROC'是用R版本4.4.2 来建造的

## Type 'citation("pROC")' for a citation.

##   
## 载入程序包：'pROC'

## The following objects are masked from 'package:stats':  
##   
## cov, smooth, var

# Function for Model Evaluation  
evaluate\_model <- function(model, data, model\_name) {  
 # Log-Likelihood  
 log\_likelihood <- logLik(model)  
 cat("Log-Likelihood for", model\_name, ":", log\_likelihood, "\n")  
   
 # Predictions  
 predictions <- predict(model, type = "response")  
 data$predicted\_classes <- ifelse(predictions > 0.5, 1, 0)  
   
 # Ensure Consistent Factor Levels  
 data$predicted\_classes <- factor(data$predicted\_classes, levels = c(0, 1))  
 data$status <- factor(data$status, levels = c(0, 1))  
   
 # Confusion Matrix  
 conf\_matrix <- confusionMatrix(data$predicted\_classes, data$status)  
 cat("Confusion Matrix for", model\_name, ":\n")  
 print(conf\_matrix)  
   
 # AUC-ROC  
 roc\_curve <- roc(as.numeric(as.character(data$status)), predictions)  
 auc\_value <- auc(roc\_curve)  
 cat("AUC for", model\_name, ":", auc\_value, "\n")  
   
 return(list(conf\_matrix = conf\_matrix, auc = auc\_value, roc\_curve = roc\_curve))  
}  
  
  
# Evaluate Models  
cat("\n### Evaluating AIC Models ###\n")

##   
## ### Evaluating AIC Models ###

aic\_forward\_results <- evaluate\_model(aic\_forward\_model, balance\_data, "AIC Forward")

## Log-Likelihood for AIC Forward : -708.6961   
## Confusion Matrix for AIC Forward :  
## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 463 230  
## 1 153 386  
##   
## Accuracy : 0.6891   
## 95% CI : (0.6624, 0.7149)  
## No Information Rate : 0.5   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.3782   
##   
## Mcnemar's Test P-Value : 0.000103   
##   
## Sensitivity : 0.7516   
## Specificity : 0.6266   
## Pos Pred Value : 0.6681   
## Neg Pred Value : 0.7161   
## Prevalence : 0.5000   
## Detection Rate : 0.3758   
## Detection Prevalence : 0.5625   
## Balanced Accuracy : 0.6891   
##   
## 'Positive' Class : 0   
##

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

## Setting direction: controls < cases

## AUC for AIC Forward : 0.7646236

aic\_backward\_results <- evaluate\_model(aic\_backward\_model, balance\_data, "AIC Backward")

## Log-Likelihood for AIC Backward : -713.853   
## Confusion Matrix for AIC Backward :  
## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 470 228  
## 1 146 388  
##   
## Accuracy : 0.6964   
## 95% CI : (0.6699, 0.722)  
## No Information Rate : 0.5   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.3929   
##   
## Mcnemar's Test P-Value : 2.809e-05   
##   
## Sensitivity : 0.7630   
## Specificity : 0.6299   
## Pos Pred Value : 0.6734   
## Neg Pred Value : 0.7266   
## Prevalence : 0.5000   
## Detection Rate : 0.3815   
## Detection Prevalence : 0.5666   
## Balanced Accuracy : 0.6964   
##   
## 'Positive' Class : 0   
##

## Setting levels: control = 0, case = 1  
## Setting direction: controls < cases

## AUC for AIC Backward : 0.7596744

aic\_both\_results <- evaluate\_model(aic\_both\_model, balance\_data, "AIC Stepwise")

## Log-Likelihood for AIC Stepwise : -713.853   
## Confusion Matrix for AIC Stepwise :  
## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 470 228  
## 1 146 388  
##   
## Accuracy : 0.6964   
## 95% CI : (0.6699, 0.722)  
## No Information Rate : 0.5   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.3929   
##   
## Mcnemar's Test P-Value : 2.809e-05   
##   
## Sensitivity : 0.7630   
## Specificity : 0.6299   
## Pos Pred Value : 0.6734   
## Neg Pred Value : 0.7266   
## Prevalence : 0.5000   
## Detection Rate : 0.3815   
## Detection Prevalence : 0.5666   
## Balanced Accuracy : 0.6964   
##   
## 'Positive' Class : 0   
##

## Setting levels: control = 0, case = 1  
## Setting direction: controls < cases

## AUC for AIC Stepwise : 0.7596744

cat("\n### Evaluating BIC Models ###\n")

##   
## ### Evaluating BIC Models ###

bic\_forward\_results <- evaluate\_model(bic\_forward\_model, balance\_data, "BIC Forward")

## Log-Likelihood for BIC Forward : -708.6961   
## Confusion Matrix for BIC Forward :  
## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 463 230  
## 1 153 386  
##   
## Accuracy : 0.6891   
## 95% CI : (0.6624, 0.7149)  
## No Information Rate : 0.5   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.3782   
##   
## Mcnemar's Test P-Value : 0.000103   
##   
## Sensitivity : 0.7516   
## Specificity : 0.6266   
## Pos Pred Value : 0.6681   
## Neg Pred Value : 0.7161   
## Prevalence : 0.5000   
## Detection Rate : 0.3758   
## Detection Prevalence : 0.5625   
## Balanced Accuracy : 0.6891   
##   
## 'Positive' Class : 0   
##

## Setting levels: control = 0, case = 1  
## Setting direction: controls < cases

## AUC for BIC Forward : 0.7646236

bic\_backward\_results <- evaluate\_model(bic\_backward\_model, balance\_data, "BIC Backward")

## Log-Likelihood for BIC Backward : -727.6008   
## Confusion Matrix for BIC Backward :  
## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 490 249  
## 1 126 367  
##   
## Accuracy : 0.6956   
## 95% CI : (0.6691, 0.7212)  
## No Information Rate : 0.5   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.3912   
##   
## Mcnemar's Test P-Value : 2.975e-10   
##   
## Sensitivity : 0.7955   
## Specificity : 0.5958   
## Pos Pred Value : 0.6631   
## Neg Pred Value : 0.7444   
## Prevalence : 0.5000   
## Detection Rate : 0.3977   
## Detection Prevalence : 0.5998   
## Balanced Accuracy : 0.6956   
##   
## 'Positive' Class : 0   
##

## Setting levels: control = 0, case = 1  
## Setting direction: controls < cases

## AUC for BIC Backward : 0.7501739

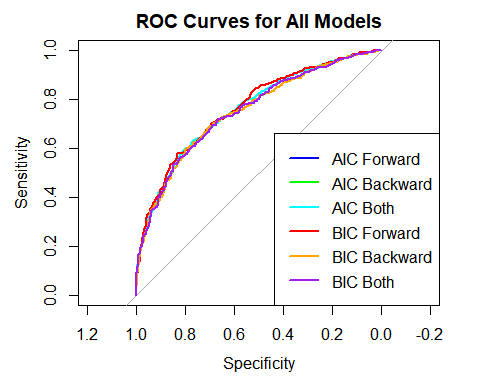
bic\_both\_results <- evaluate\_model(bic\_both\_model, balance\_data, "BICS tepwise")

## Log-Likelihood for BICS tepwise : -723.3745   
## Confusion Matrix for BICS tepwise :  
## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 481 250  
## 1 135 366  
##   
## Accuracy : 0.6875   
## 95% CI : (0.6608, 0.7133)  
## No Information Rate : 0.5   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.375   
##   
## Mcnemar's Test P-Value : 6.248e-09   
##   
## Sensitivity : 0.7808   
## Specificity : 0.5942   
## Pos Pred Value : 0.6580   
## Neg Pred Value : 0.7305   
## Prevalence : 0.5000   
## Detection Rate : 0.3904   
## Detection Prevalence : 0.5933   
## Balanced Accuracy : 0.6875   
##   
## 'Positive' Class : 0   
##

## Setting levels: control = 0, case = 1  
## Setting direction: controls < cases

## AUC for BICS tepwise : 0.7510436

# Plot ROC Curves for All Models  
plot(aic\_forward\_results$roc\_curve, col = "blue", lwd = 2, main = "ROC Curves for All Models")  
lines(aic\_backward\_results$roc\_curve, col = "green", lwd = 2)  
lines(aic\_both\_results$roc\_curve, col = "cyan", lwd = 2)  
lines(bic\_forward\_results$roc\_curve, col = "red", lwd = 2)  
lines(bic\_backward\_results$roc\_curve, col = "orange", lwd = 2)  
lines(bic\_both\_results$roc\_curve, col = "purple", lwd = 2)  
  
legend("bottomright",   
 legend = c("AIC Forward", "AIC Backward", "AIC Both",   
 "BIC Forward", "BIC Backward", "BIC Both"),  
 col = c("blue", "green", "cyan", "red", "orange", "purple"),   
 lwd = 2)



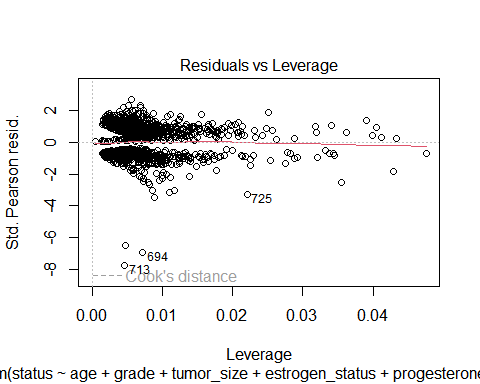
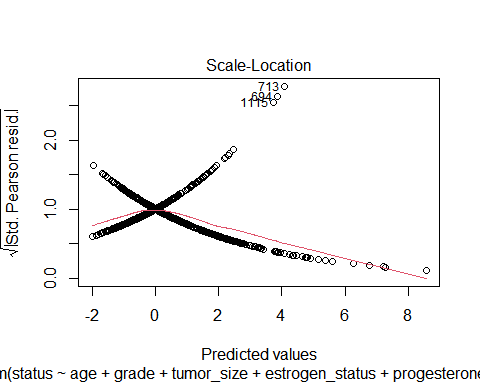
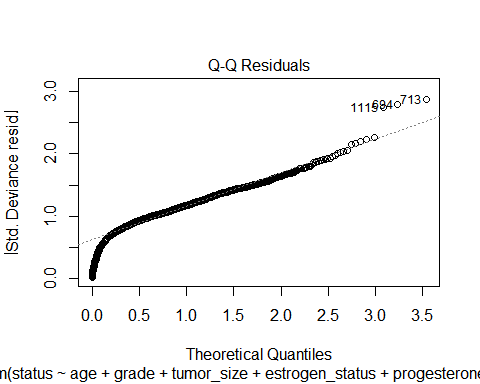
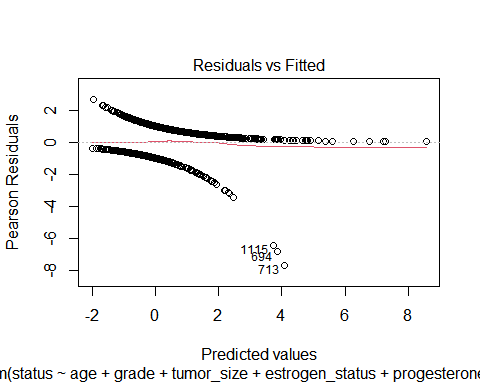
summary(bic\_forward\_model)

##   
## Call:  
## glm(formula = status ~ (age + grade + tumor\_size + estrogen\_status +   
## progesterone\_status + regional\_node\_examined + reginol\_node\_positive)^2 +   
## +marital\_status\_Single + marital\_status\_Widowed + marital\_status\_Separated +   
## x6th\_stage\_IIB + x6th\_stage\_IIIA + x6th\_stage\_IIIB + x6th\_stage\_IIIC +   
## a\_stage\_regional, family = binomial(), data = balance\_data\_dummies)  
##   
## Coefficients:  
## Estimate Std. Error z value  
## (Intercept) -5.0315671 2.7764165 -1.812  
## age 0.0612314 0.0422038 1.451  
## grade 0.8686971 0.8222391 1.057  
## tumor\_size 0.0480101 0.0301414 1.593  
## estrogen\_status 0.7586018 2.3413419 0.324  
## progesterone\_status -1.6544489 1.8996219 -0.871  
## regional\_node\_examined 0.0313134 0.0822064 0.381  
## reginol\_node\_positive 0.2185913 0.1343681 1.627  
## marital\_status\_Single 0.2066461 0.1787612 1.156  
## marital\_status\_Widowed 0.2325247 0.2630843 0.884  
## marital\_status\_Separated 0.8567306 0.5595292 1.531  
## x6th\_stage\_IIB 0.5730647 0.1991393 2.878  
## x6th\_stage\_IIIA 0.7328995 0.2651780 2.764  
## x6th\_stage\_IIIB 1.6098447 0.5854185 2.750  
## x6th\_stage\_IIIC 1.0192877 0.4531758 2.249  
## a\_stage\_regional -0.3363202 0.5100840 -0.659  
## age:grade -0.0100271 0.0119662 -0.838  
## age:tumor\_size -0.0002938 0.0003529 -0.833  
## age:estrogen\_status 0.0054306 0.0318613 0.170  
## age:progesterone\_status 0.0078672 0.0230871 0.341  
## age:regional\_node\_examined -0.0007101 0.0010501 -0.676  
## age:reginol\_node\_positive -0.0008303 0.0016078 -0.516  
## grade:tumor\_size 0.0030330 0.0055557 0.546  
## grade:estrogen\_status -0.0289540 0.4897073 -0.059  
## grade:progesterone\_status 0.1601791 0.3183625 0.503  
## grade:regional\_node\_examined -0.0033850 0.0144429 -0.234  
## grade:reginol\_node\_positive -0.0145921 0.0220955 -0.660  
## tumor\_size:estrogen\_status -0.0513603 0.0176299 -2.913  
## tumor\_size:progesterone\_status 0.0096456 0.0096120 1.003  
## tumor\_size:regional\_node\_examined -0.0001558 0.0004738 -0.329  
## tumor\_size:reginol\_node\_positive 0.0010221 0.0007258 1.408  
## estrogen\_status:progesterone\_status 0.2201728 0.8695034 0.253  
## estrogen\_status:regional\_node\_examined 0.0040641 0.0389264 0.104  
## estrogen\_status:reginol\_node\_positive 0.0327185 0.0735757 0.445  
## progesterone\_status:regional\_node\_examined -0.0062403 0.0258520 -0.241  
## progesterone\_status:reginol\_node\_positive -0.0627842 0.0551059 -1.139  
## regional\_node\_examined:reginol\_node\_positive -0.0023211 0.0013192 -1.760  
## Pr(>|z|)   
## (Intercept) 0.06995 .   
## age 0.14682   
## grade 0.29074   
## tumor\_size 0.11120   
## estrogen\_status 0.74594   
## progesterone\_status 0.38379   
## regional\_node\_examined 0.70327   
## reginol\_node\_positive 0.10378   
## marital\_status\_Single 0.24769   
## marital\_status\_Widowed 0.37678   
## marital\_status\_Separated 0.12573   
## x6th\_stage\_IIB 0.00401 \*\*  
## x6th\_stage\_IIIA 0.00571 \*\*  
## x6th\_stage\_IIIB 0.00596 \*\*  
## x6th\_stage\_IIIC 0.02450 \*   
## a\_stage\_regional 0.50968   
## age:grade 0.40206   
## age:tumor\_size 0.40513   
## age:estrogen\_status 0.86466   
## age:progesterone\_status 0.73328   
## age:regional\_node\_examined 0.49893   
## age:reginol\_node\_positive 0.60554   
## grade:tumor\_size 0.58512   
## grade:estrogen\_status 0.95285   
## grade:progesterone\_status 0.61487   
## grade:regional\_node\_examined 0.81470   
## grade:reginol\_node\_positive 0.50899   
## tumor\_size:estrogen\_status 0.00358 \*\*  
## tumor\_size:progesterone\_status 0.31562   
## tumor\_size:regional\_node\_examined 0.74231   
## tumor\_size:reginol\_node\_positive 0.15908   
## estrogen\_status:progesterone\_status 0.80010   
## estrogen\_status:regional\_node\_examined 0.91685   
## estrogen\_status:reginol\_node\_positive 0.65654   
## progesterone\_status:regional\_node\_examined 0.80926   
## progesterone\_status:reginol\_node\_positive 0.25456   
## regional\_node\_examined:reginol\_node\_positive 0.07849 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1707.9 on 1231 degrees of freedom  
## Residual deviance: 1417.4 on 1195 degrees of freedom  
## AIC: 1491.4  
##   
## Number of Fisher Scoring iterations: 6

summary(bic\_backward\_model)

##   
## Call:  
## glm(formula = status ~ age + grade + tumor\_size + estrogen\_status +   
## progesterone\_status + regional\_node\_examined + reginol\_node\_positive +   
## tumor\_size:estrogen\_status, family = binomial(), data = balance\_data\_dummies)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.386362 0.696175 -4.864 1.15e-06 \*\*\*  
## age 0.024875 0.007016 3.546 0.000392 \*\*\*  
## grade 0.410135 0.103654 3.957 7.60e-05 \*\*\*  
## tumor\_size 0.058855 0.015565 3.781 0.000156 \*\*\*  
## estrogen\_status 1.110410 0.514671 2.158 0.030966 \*   
## progesterone\_status -0.640839 0.178206 -3.596 0.000323 \*\*\*  
## regional\_node\_examined -0.032673 0.009024 -3.621 0.000294 \*\*\*  
## reginol\_node\_positive 0.135035 0.014992 9.007 < 2e-16 \*\*\*  
## tumor\_size:estrogen\_status -0.051505 0.015888 -3.242 0.001188 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1707.9 on 1231 degrees of freedom  
## Residual deviance: 1455.2 on 1223 degrees of freedom  
## AIC: 1473.2  
##   
## Number of Fisher Scoring iterations: 6

plot(bic\_backward\_model)



Evaluate the performance of your modells). ls your model achieving similar performancebetween the majority race group “White” and the minority “Black” (or “Black” + “Other”)? lfnot, could you try to improve the fairness (i.e., reducing the gap of prediction performancebetween the majority and minority) of your model(s)?

library(caret)  
library(pROC)  
  
# Assuming 'race' is your race variable and 'model' is your trained model  
evaluate\_performance <- function(data, model, group\_name) {  
 # Subset data  
 subset\_data <- data[data$race == group\_name, ]  
   
 # Predictions  
 predictions <- predict(bic\_backward\_model, newdata = subset\_data, type = "response")  
 predicted\_classes <- ifelse(predictions > 0.5, 1, 0)  
   
 # Confusion Matrix  
 conf\_matrix <- confusionMatrix(factor(predicted\_classes), factor(subset\_data$status))  
 print(paste("Confusion Matrix for", group\_name))  
 print(conf\_matrix)  
   
 # AUC-ROC  
 roc\_curve <- roc(subset\_data$status, predictions)  
 auc\_value <- auc(roc\_curve)  
 print(paste("AUC for", group\_name, ":", auc\_value))  
}  
  
# Evaluate for 'White' group  
evaluate\_performance(balance\_data, bic\_backward\_model, "White")

## [1] "Confusion Matrix for White"  
## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 421 202  
## 1 104 308  
##   
## Accuracy : 0.7043   
## 95% CI : (0.6755, 0.732)  
## No Information Rate : 0.5072   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.4069   
##   
## Mcnemar's Test P-Value : 2.938e-08   
##   
## Sensitivity : 0.8019   
## Specificity : 0.6039   
## Pos Pred Value : 0.6758   
## Neg Pred Value : 0.7476   
## Prevalence : 0.5072   
## Detection Rate : 0.4068   
## Detection Prevalence : 0.6019   
## Balanced Accuracy : 0.7029   
##   
## 'Positive' Class : 0   
##

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

## Setting direction: controls < cases

## [1] "AUC for White : 0.757971988795518"

# Evaluate for 'Black' group  
evaluate\_performance(balance\_data, bic\_backward\_model, "Black")

## [1] "Confusion Matrix for Black"  
## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 33 27  
## 1 12 46  
##   
## Accuracy : 0.6695   
## 95% CI : (0.5769, 0.7533)  
## No Information Rate : 0.6186   
## P-Value [Acc > NIR] : 0.14843   
##   
## Kappa : 0.3416   
##   
## Mcnemar's Test P-Value : 0.02497   
##   
## Sensitivity : 0.7333   
## Specificity : 0.6301   
## Pos Pred Value : 0.5500   
## Neg Pred Value : 0.7931   
## Prevalence : 0.3814   
## Detection Rate : 0.2797   
## Detection Prevalence : 0.5085   
## Balanced Accuracy : 0.6817   
##   
## 'Positive' Class : 0   
##

## Setting levels: control = 0, case = 1  
## Setting direction: controls < cases

## [1] "AUC for Black : 0.744292237442922"

# Evaluate for 'Black' + 'Other' group  
evaluate\_performance(balance\_data, bic\_backward\_model, c("Black", "Other"))

## [1] "Confusion Matrix for Black" "Confusion Matrix for Other"  
## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 35 21  
## 1 8 27  
##   
## Accuracy : 0.6813   
## 95% CI : (0.5753, 0.7751)  
## No Information Rate : 0.5275   
## P-Value [Acc > NIR] : 0.002066   
##   
## Kappa : 0.3706   
##   
## Mcnemar's Test P-Value : 0.025858   
##   
## Sensitivity : 0.8140   
## Specificity : 0.5625   
## Pos Pred Value : 0.6250   
## Neg Pred Value : 0.7714   
## Prevalence : 0.4725   
## Detection Rate : 0.3846   
## Detection Prevalence : 0.6154   
## Balanced Accuracy : 0.6882   
##   
## 'Positive' Class : 0   
##

## Setting levels: control = 0, case = 1  
## Setting direction: controls < cases

## [1] "AUC for Black : 0.743217054263566" "AUC for Other : 0.743217054263566"