Code ▾

Multivariable Logistic Regression for Breast Cancer Recurrence Prediction

reference: https://statsandr.com/blog/binary-logistic-regression-in-r/ (https://statsandr.com/blog/binary-logistic-regression-in-r/)

1. Load Preprocessed Breast Cancer Dataset

Hide

```
breast_data = read.csv("breast_cancer_new.csv", row.names = 1)
```

2. Factorize and relevel values of some variables

Hide

```
breast_data$class = factor(breast_data$class)
breast_data$menopause = factor(breast_data$menopause)
breast_data$node.caps = factor(breast_data$node.caps)
breast_data$breast = factor(breast_data$breast)
breast_data$breast.quad = factor(breast_data$breast.quad)
breast_data$irradiation = factor(breast_data$irradiation)
breast_data$age = factor(breast_data$age,
                         levels = c("20-29", "30-39" ,"40-49","50-59" ,"60-69",
                                    "70-79"))
breast_data$inv.nodes = factor(breast_data$inv.nodes,
                         levels = c("0-2", "3-5" ,"6-8","9-11" ,"12-14",
                                    "15-17","24-26"))
breast_data$deg.malignancy = factor(breast_data$deg.malignancy,
                                    levels = c("1","2","3"))
breast_data$tumor.size = factor(breast_data$tumor.size,
                                           levels = c("0-4","5-9", "10-14","15-19", "20-24","25-
29" ,"30-34" ,"35-39" ,"40-44" ,"45-49","50-54" ))
```

Hide

```
summary(breast_data)
```

```
                  class        age       menopause     tumor.size inv.nodes 
 no-recurrence-events:196   20-29: 1   ge40   :123   30-34  :57   0-2  :209  
 recurrence-events   : 81   30-39:36   lt40   :  5   25-29  :51   3-5  : 34  
                            40-49:89   premeno:149   20-24  :48   6-8  : 17  
                            50-59:91                 15-19  :29   9-11 :  7  
                            60-69:55                 10-14  :28   12-14:  3  
                            70-79: 5                 40-44  :22   15-17:  6  
                                                     (Other):42   24-26:  1  
 node.caps deg.malignancy   breast       breast.quad  irradiation
 no :221   1: 66          left :145   central  : 21   no :215    
 yes: 56   2:129          right:132   left_low :106   yes: 62    
           3: 82                      left_up  : 94              
                                      right_low: 23              
                                      right_up : 33              
```

3. Relabel the outcome

Hide

```
breast_data$class <- ifelse(breast_data$class == 
                                   "recurrence-events", 1,0)
```

4. Build Logistic Regression Model with Significantly Associated Variables from Fisher’s Exact Test

Hide

```
full_model = glm(class~tumor.size+inv.nodes+node.caps+deg.malignancy+irradiation,
            data = breast_data, family = binomial)
summary(full_model)
```

```

Call:
glm(formula = class ~ tumor.size + inv.nodes + node.caps + deg.malignancy + 
    irradiation, family = binomial, data = breast_data)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)  
(Intercept)       -2.09396    1.12375  -1.863   0.0624 .
tumor.size5-9    -14.55880 1188.56199  -0.012   0.9902  
tumor.size10-14   -1.64782    1.52972  -1.077   0.2814  
tumor.size15-19    0.23000    1.20319   0.191   0.8484  
tumor.size20-24    0.41668    1.15208   0.362   0.7176  
tumor.size25-29    0.67633    1.14340   0.592   0.5542  
tumor.size30-34    0.77383    1.13540   0.682   0.4955  
tumor.size35-39    0.31513    1.23817   0.255   0.7991  
tumor.size40-44   -0.03074    1.23667  -0.025   0.9802  
tumor.size45-49    0.19375    1.76154   0.110   0.9124  
tumor.size50-54    1.05659    1.34419   0.786   0.4318  
inv.nodes3-5       0.91760    0.51328   1.788   0.0738 .
inv.nodes6-8       0.99676    0.67694   1.472   0.1409  
inv.nodes9-11      1.58372    0.98914   1.601   0.1094  
inv.nodes12-14     0.41038    1.33457   0.308   0.7585  
inv.nodes15-17     0.49179    0.95309   0.516   0.6059  
inv.nodes24-26    16.33783 2399.54482   0.007   0.9946  
node.capsyes       0.21413    0.47163   0.454   0.6498  
deg.malignancy2   -0.13564    0.46691  -0.291   0.7714  
deg.malignancy3    1.16160    0.46726   2.486   0.0129 *
irradiationyes     0.52979    0.36817   1.439   0.1501  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 334.78  on 276  degrees of freedom
Residual deviance: 268.68  on 256  degrees of freedom
AIC: 310.68

Number of Fisher Scoring iterations: 15
```

5. Perform stepwise variable selection to improve the AIC

Hide

```
stepwise_selection = step(full_model, direction = "both", criterion = "AIC")
```

```
Start:  AIC=310.68
class ~ tumor.size + inv.nodes + node.caps + deg.malignancy + 
    irradiation

                 Df Deviance    AIC
- tumor.size     10   280.33 302.33
- inv.nodes       6   274.51 304.51
- node.caps       1   268.88 308.89
<none>                268.68 310.68
- irradiation     1   270.73 310.73
- deg.malignancy  2   284.31 322.31

Step:  AIC=302.33
class ~ inv.nodes + node.caps + deg.malignancy + irradiation

                 Df Deviance    AIC
- inv.nodes       6   287.00 297.00
- node.caps       1   280.63 300.63
- irradiation     1   282.27 302.27
<none>                280.33 302.33
+ tumor.size     10   268.68 310.68
- deg.malignancy  2   299.44 317.44

Step:  AIC=297
class ~ node.caps + deg.malignancy + irradiation

                 Df Deviance    AIC
<none>                287.00 297.00
- irradiation     1   290.72 298.72
- node.caps       1   293.44 301.44
+ inv.nodes       6   280.33 302.33
+ tumor.size     10   274.51 304.51
- deg.malignancy  2   308.12 314.12
```

Hide

```
# print selected variables
selected_variables <- attr(terms(stepwise_selection), "term.labels")
selected_variables
```

```
[1] "node.caps"      "deg.malignancy" "irradiation"
```

6. Build logistic regression model with selected variables and without interaction

Hide

```
model_selectedVar = glm(class~node.caps+deg.malignancy+irradiation,
                        data = breast_data, family = "binomial")
summary(model_selectedVar)
```

```


Call:
glm(formula = class ~ node.caps + deg.malignancy + irradiation,
    family = "binomial", data = breast_data)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.8956     0.3608  -5.254 1.49e-07 ***
node.capsyes      0.9005     0.3544   2.541 0.011053 *
deg.malignancy2   0.1971     0.4373   0.451 0.652215
deg.malignancy3   1.5156     0.4441   3.413 0.000643 ***
irradiationyes    0.6617     0.3405   1.943 0.051982 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 334.78  on 276  degrees of freedom
Residual deviance: 287.00  on 272  degrees of freedom
AIC: 297

Number of Fisher Scoring iterations: 4
```

Hide

```
# check the effect of interactions between the variables
model_node_deg_inter = glm(class~node.caps*deg.malignancy+irradiation,
                        data = breast_data, family = binomial)
model_node_irrad_inter = glm(class~node.caps*irradiation+deg.malignancy,
                        data = breast_data, family = binomial)
model_deg_irrad_inter = glm(class~node.caps+deg.malignancy*irradiation,
                        data = breast_data, family = binomial)
```

Hide

```
# tabulate the AIC of the models with interaction and without interaction
data.frame(model = c("model_WO_inter","node_deg", "node_irrad","deg_irrad"),
           AIC = c(model_selectedVar$aic,model_node_deg_inter$aic,
                   model_node_irrad_inter$aic,model_deg_irrad_inter$aic))
```

model <chr>	AIC <dbl>
model_WO_inter	297.0034
node_deg	296.5151

model <chr>	AIC <dbl>
node_irrad	298.2297
deg_irrad	298.2884

4 rows

7. Perform likelihood ratio test to check the effect of each variable by omitting it in a new model and comparing it with model with interaction between node.caps and deg.malignancy

Hide

```
# check the effect of deg.malignancy variable
degMalignancy_omit = glm(class~node.caps+irradiation,data = breast_data,
                         family = binomial)
summary(degMalignancy_omit)
```

```

Call:
glm(formula = class ~ node.caps + irradiation, family = binomial,
    data = breast_data)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.3643     0.1750  -7.795 6.43e-15 ***
node.capsyes     1.2120     0.3303   3.670 0.000243 ***
irradiationyes   0.7499     0.3243   2.312 0.020767 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 334.78  on 276  degrees of freedom
Residual deviance: 308.12  on 274  degrees of freedom
AIC: 314.12

Number of Fisher Scoring iterations: 4
```

Hide

```
anova(degMalignancy_omit, model_node_deg_inter, test = "LRT")
```

```
Analysis of Deviance Table

Model 1: class ~ node.caps + irradiation
Model 2: class ~ node.caps * deg.malignancy + irradiation
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)    
1       274     308.12                          
2       271     284.51  3   23.609 3.014e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hide

```
# check the effect of irradiation variable
irradiation_omit = glm(class~node.caps*deg.malignancy,
                       data = breast_data, family = binomial)
summary(irradiation_omit)
```

```


Call:
glm(formula = class ~ node.caps * deg.malignancy, family = binomial, 
    data = breast_data)

Coefficients: (1 not defined because of singularities)
                             Estimate Std. Error z value Pr(>|z|)    
(Intercept)                   -1.8458     0.3587  -5.146 2.66e-07 ***
node.capsyes                   1.5790     0.5160   3.060  0.00221 ** 
deg.malignancy2                0.4227     0.4367   0.968  0.33305    
deg.malignancy3                1.4564     0.4567   3.189  0.00143 ** 
node.capsyes:deg.malignancy2  -0.9669     0.7133  -1.355  0.17527    
node.capsyes:deg.malignancy3       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: 334.78  on 276  degrees of freedom
Residual deviance: 288.82  on 272  degrees of freedom
AIC: 298.82

Number of Fisher Scoring iterations: 4
```

Hide

```
anova(irradiation_omit, model_node_deg_inter, test = "LRT")
```

```
Analysis of Deviance Table

Model 1: class ~ node.caps * deg.malignancy
Model 2: class ~ node.caps * deg.malignancy + irradiation
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       272     288.82
2       271     284.51  1   4.3039  0.03802 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hide

```
# check the effect of node.caps variable
nodeCaps_omit = glm(class~deg.malignancy+irradiation,data = breast_data,
                         family = binomial)
summary(nodeCaps_omit)
```

```


Call:
glm(formula = class ~ deg.malignancy + irradiation, family = binomial,
    data = breast_data)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.9166     0.3617  -5.299 1.17e-07 ***
deg.malignancy2   0.3822     0.4269   0.895  0.37055
deg.malignancy3   1.7832     0.4302   4.145 3.40e-05 ***
irradiationyes    0.8824     0.3244   2.720  0.00652 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 334.78  on 276  degrees of freedom
Residual deviance: 293.45  on 273  degrees of freedom
AIC: 301.45

Number of Fisher Scoring iterations: 4
```

Hide

```
anova(nodeCaps_omit, model_node_deg_inter, test = "LRT")
```

```
Analysis of Deviance Table

Model 1: class ~ deg.malignancy + irradiation
Model 2: class ~ node.caps * deg.malignancy + irradiation
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       273     293.44
2       271     284.51  2     8.93   0.0115 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpretation: At 5% signifincant level, the model with selected variables and interaction between node.caps and deg.malignancy variables is significantly different from that of without each of the variables. This means the selected variables and interaction between node.caps and deg.malignancy are significantly associated with breast cancer recurrence.

8. Evaluate final logistic regression model with selected variables and interaction

Hide

```
library(ResourceSelection)
```

```
ResourceSelection 0.3-6      2023-06-27
```

Hide

```
# build final logistic regression model
final_model = glm(class~node.caps*deg.malignancy+irradiation,
                        data = breast_data, family = binomial)
```

Hide

```
summary(final_model)
```

```

Call:
glm(formula = class ~ node.caps * deg.malignancy + irradiation, 
    family = binomial, data = breast_data)

Coefficients: (1 not defined because of singularities)
                             Estimate Std. Error z value Pr(>|z|)    
(Intercept)                   -1.9014     0.3611  -5.265  1.4e-07 ***
node.capsyes                   1.4637     0.5232   2.798  0.00515 ** 
deg.malignancy2                0.3364     0.4409   0.763  0.44545    
deg.malignancy3                1.3239     0.4634   2.857  0.00428 ** 
irradiationyes                 0.7253     0.3469   2.091  0.03654 *  
node.capsyes:deg.malignancy2  -1.1258     0.7281  -1.546  0.12208    
node.capsyes:deg.malignancy3       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: 334.78  on 276  degrees of freedom
Residual deviance: 284.52  on 271  degrees of freedom
AIC: 296.52

Number of Fisher Scoring iterations: 4
```

Hide

```
# perform Hosmer-Lemeshow Test to check the goodness-of-fit of the model
hoslem.test(as.numeric(breast_data$class), fitted(final_model), g = 2)
```

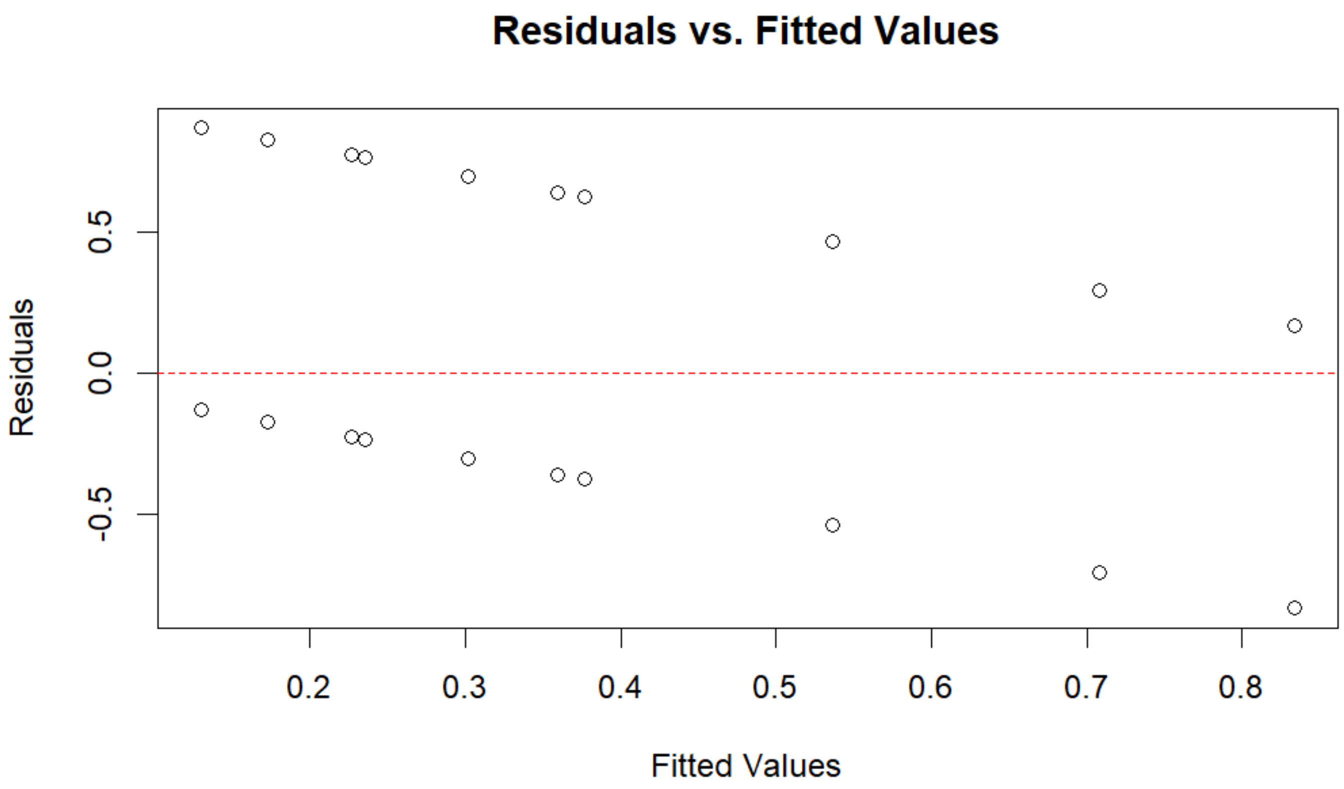
```


    Hosmer and Lemeshow goodness of fit (GOF) test

data:  as.numeric(breast_data$class), fitted(final_model)
X-squared = 0.00041064, df = 0, p-value < 2.2e-16
```

Hide

```
# Residual vs Fitted Values Analysis
model_residuals = residuals(final_model, type = "response")
plot(fitted(final_model), model_residuals,
     xlab = "Fitted Values",
     ylab = "Residuals",
     main = "Residuals vs. Fitted Values")
abline(h = 0, col = "red", lty = 2)
```



The selected model has a poor fit based on Hosmer-Lemeshow test and residuals vs fitted values analysis. Further improvement of the model is necessary.