Breast Cancer

FINAL COURSE PROJECT
03/12/2023

Alba Mas and Maria Cobo

Index

1. Context and goals

Nowadays, breast cancer is one of the most common cancers in women (other than skin cancer) in which abnormal breast cells grow out of control and form tumors (*Breast Cancer*, 2023). If left unchecked, it can spread throughout the body like to the lymph nodes under the arms.

Moreover, about 1 in 7 women are diagnosed with breast cancer during their lifetime and if detected at an early age, there is a good chance of recovery (*Overview - - - Breast Cancer in Women*, n.d.). So, it is vital that women check their breasts regularly.

However, even though someone recovers from breast cancer, there is a chance of recurrence, the cancer could come back. Therefore, the goal of our investigation is to find out which variables will increase the chances of recurrence in breast cancer. To do so, we will build an efficient model that predicts the individual's probability of recurrence of breast cancer and use as many significant predictors as possible. To create this model, we will be using a data set that has been taken from the University Medical Centre Institute of Oncology, Ljubljana, Yugoslavia (M. Zwitter and M. Soklic).

This data set includes 9 observations for 286 different individuals. The following observations:

Variable Name	**Description**	**Type**	**Values**
Class	Dependent variable that indicates if there is or not recurrence	Binary	no-recurrence-events, recurrence-events
Age	Range of individual's age	Categorical	10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99
Menopause	Menopause status of the individual	Categorical	1 = lt40 2 = ge40 (40 and older) 3 = premeno
Tumor-size	Size of the tumor, specified in range	Categorical	0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59
Inv-nodes	Number of involved lymph nodes within a range	Categorical	0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39
Node-caps	Presence or absence of node capsulation	Binary	yes = 1 no = 2
Deg-malig	Degree of malignancy of the tumor	Integer	1, 2, 3
Breast	Breast affected	Binary	1 = left, 2 = right
Breast-quad	Quadrant of the breast affected	Categorical	2 = left-up, 3 = left-low, 4 = right-up, 5 = right-low, 6 = central
Irradiat	Whether the patient has received radiation treatment or not	Binary	1 =yes 2 = no

1.2 Expectations on the results

Before performing any computations, we have done some research on the factors that increase the risk in recurrence of breast cancer. According to the Mayo Clinic (*Recurrent Breast Cancer - Symptoms and Causes*, 2022), one of the factors that may risk recurrence of breast cancer is the lymph node involvement. If the cancer was found nearby lymph nodes this increases the risk of recurrence. Moreover, another factor is the size of the tumor. If the individual had a larger tumor size, this will also increase the risk of recurrence. Also, those that don't undertake radiation treatment and those under the age of 35 will also increase the risk of recurrence.

2. Description of the dataset

To perform this study, we are going to analyze a breast cancer dataset, using the Class (recurrence or no-recurrence events) variable as response variable and the other variables as predictors to assess if there is any significance or relation between them. We expect certain attributes to have clinically meaningful distributions that align with known patterns in breast cancer. In this study, we have 8 predictor variables (Age, Menopause, Tumor-size, Inv-nodes, Node-caps, Deg-malig, Breast, Breast-quad and Irradiat) and one response variable (Class).

Additionally, our data set had some missing values so we have removed them. We changed the '?' values by NA and then omitted these. Now, instead of 286 individuals we have 277.

2.1 Descriptive statistics

Age range	No-recurrence	Recurrence
20-29	1	0
30-39	21	15
40-49	62	27
50-59	69	22
60-69	38	17
70-79	55	0

In order to do a descriptive analysis, we shall do a summary on our data to quantify our response variable (Class). In our data set, after removing the missing values, there are 196 individuals that will have no recurrence and 81 individuals with recurrence of breast cancer.

We can see that breast cancer in younger and older ages is less common. However, the highest odds of recurrence and no-recurrence is between the ages "40-69". Especially, age "40-49" that has the highest value on recurrence events while ages "50-59" has the highest values in no-recurrence events.

2.3 Plot the data

Since we are interested in identifying significant risk factors associated with breast cancer recurrence, we will start by plotting the Class distribution and observing in a graphical way the difference between recurrence and no-recurrence events. We can observe that there are a lot more cases of no-recurrence in our dataset.

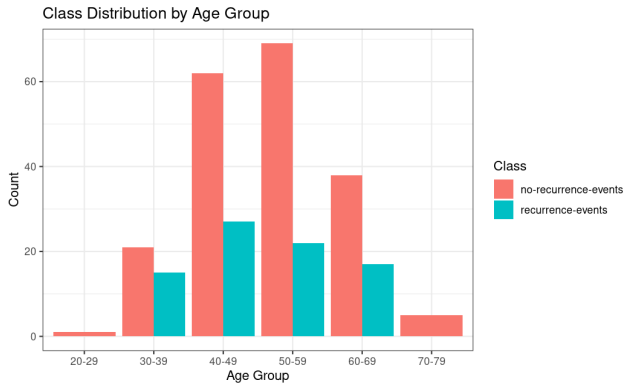
2.4. Data Analysis

To visualize the relationship between different predictor variables and the Class response variable (recurrence and no-recurrence events), we will analyze every predictor separately and then with these results we will build a model with as many predictors as possible, to predict the likelihood of a person having recurrence or not of breast cancer.

AGE:

We can start by performing a plot to visualize how the distribution of class (recurrence and no-recurrence events) varies across different age groups. First, we can generate a plot of the individuals in each age group and if they have recurrence or not and, we can observe that this plot seems to follow a normal distribution. Here, we can see the same results as we saw in the point 2.1, the age group "40-49" has the higher number of individuals with recurrence.

Moreover, we perform a linear regression model with a p-value of 0.05 and we observe that there isn't a very strong relationship (>0.05) between age and the recurrence of breast cancer. So, we can accept the Null hypothesis and state that there is not enough evidence to suggest that 'Age' significantly predicts the Class variable in the model.




```
Call:
glm(formula = Class ~ as.numeric(Age), family = binomial(link = "logit"),
    data = breast_cancer)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.2935     0.4823  -0.609    0.543
as.numeric(Age)  -0.1561     0.1290  -1.210    0.226

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 348.05  on 285  degrees of freedom
Residual deviance: 346.57  on 284  degrees of freedom
AIC: 350.57

Number of Fisher Scoring iterations: 4
```

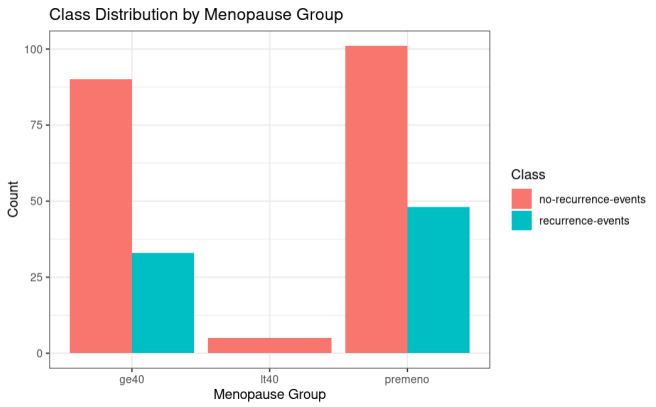
Additionally, if we look at the coefficients of this model, we can compute the estimated logit:

$$logit(Age) = -0.2935 - 0.1561 \times Age$$

And, we observe that for every one unit increase in the age value, the logit function decreases by 0.8260866. As it is a negative value, there is a negative correlation between the two variables (age and class), which means that as the age increases, the probability for recurrence in breast cancer decreases.

MENOPAUSE:

Nextly, we are going to look into the menopause predictor. First, we compute a plot of the menopause status and whether there is recurrence or not. In this plot we can see that the highest recurrence and no-recurrence is in the premenopausal group. Then, to test if this is a significant predictor to recurrence we perform a linear regression model.




```
Call:
glm(formula = Class ~ as.numeric(Menopause), family = binomial(link = "logit"),
    data = breast_cancer)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         -1.1671     0.3189  -3.660 0.000252 ***
as.numeric(Menopause)   0.1336     0.1351   0.989 0.322498
---
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: 333.80  on 275  degrees of freedom
AIC: 337.8

Number of Fisher Scoring iterations: 4
```

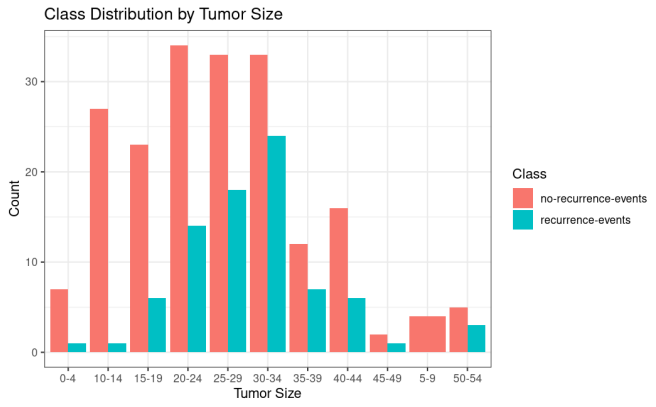
Our predictor variable ‘Menopause’ is categorical (lt40, ge40 and premeno) and when converted into a numerical value, one category becomes the reference level against which other categories are compared. In our model results, the Intercept value is less than 0.05 so, when Menopause is at its reference level, it significantly affects the recurrence (Class). However, the coefficient of Menopause is not statistically significant (p-value > 0.05) suggesting that the numerical value of Menopause is not statistically significant. There is no strong evidence to say that Menopause has a significant impact on the likelihood of cancer recurrence and therefore, we accept the null hypothesis.

Additionally, our estimated logit function is:

$$logit(Menopause) = -1.1671 + 0.1336 \times Menopause$$

The estimated logit function suggests that for each unit increase in the Class value, the log-odds increase by 1.142949 and as the coefficient of Menopause is positive, there is a positive correlation between the two variables.

TUMOR SIZE:




```
Call:
glm(formula = Class ~ as.numeric(Tumor_Size), family = binomial(link = "logit"),
    data = breast_cancer)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -1.63445    0.35112  -4.655 3.24e-06 ***
as.numeric(Tumor_Size)  0.14430    0.06096   2.367   0.0179 *
---
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: 329.10  on 275  degrees of freedom
AIC: 333.1

Number of Fisher Scoring iterations: 4
```

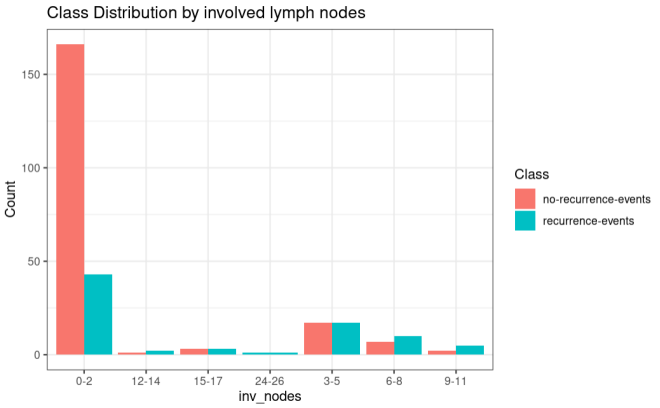
We believe that the tumor size in this dataset is set to be in millimeters (mm). Therefore, we can see that tumors with sizes between 2 and 4 cm increase the recurrence of breast cancer. According to a paper in the NCBI, a tumor size larger than 2 cm, have been shown to increase the risk of death after breast cancer diagnosis and the risk of locoregional recurrence and metastases. (*Factors Associated*

With Breast Cancer Recurrences or Mortality and Dynamic Prediction of Death Using History of Cancer Recurrences: The French E3N Cohort, 2018).
Moreover, in this plot we can see that the recurrence events seem to follow a normal distribution but it is not the case for the no-recurrence events.

In our model results, we can observe a p-value lower than 0.05 meaning that there is enough evidence to say that the impact of the tumor size on the recurrence is significant. In this case, we can reject the null hypothesis and say that the tumor size has a significant impact on the likelihood of cancer recurrence and the logit function is:
$logit(\text{Tumor\_Size}) = -1.63445 + 0.14430 \times \text{Tumor\_Size}$

Also, to continue to investigate the coefficients, we can compute the odds-log ratio which is 1.155232. This implies that for each one-unit increase in the tumor size, the odds of the recurrence will increase by 1.155232. The positive sign in the coefficient indicates a positive relationship between tumor size and recurrence, as one increases so does the other.

INVOLVED LYMPH NODES:



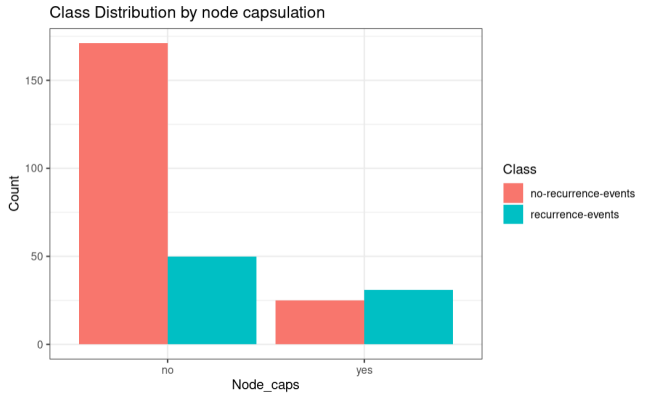

```
Call:
glm(formula = Class ~ as.numeric(inv.nodes), family = binomial(link = "logit"),
    data = breast_cancer)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         -1.66095    0.21179  -7.842 4.42e-15 ***
as.numeric(inv.nodes)  0.35288    0.06891   5.121 3.04e-07 ***
---
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: 307.49  on 275  degrees of freedom
AIC: 311.49

Number of Fisher Scoring iterations: 4
```

The involvement of lymph nodes is a high risk factor on recurrence of breast cancer. In our plot, we can see that the highest value of recurrence and non recurrence events occur in the smallest range of lymph nodes (0-2). On one hand, in case of recurrence, we believe that this could be happening due to tumors in fewer lymph nodes being more invasive and aggressive that can spread rapidly without affecting a larger number of lymph nodes. On the other hand, cancer in fewer lymph nodes could lead to undetected cancer leading to higher chances of recurrence.

Nextly, our model shows both in the intercept and the coefficient a high significance (p-value < 0.05) on the impact on the likelihood of ‘Class’. Meaning that the number of lymph nodes has a significant impact on the recurrence of breast cancer. Additionally, the logit function is:
$logit(inv.nodes) = -1.66095 + 0.35288 \times inv.nodes$

The positive sign in the coefficient means that as the number of lymph nodes increases, the log-odds of recurrence also increases. Additionally, the odd ratio test is 1.423155. For every one-unit increase in the ‘inv.nodes’ , the odds of the recurrence will increase by 1.423155.

NODE ENCAPSULATION:




```
Call:
glm(formula = Class ~ as.numeric(node.caps), family = binomial(link = "logit"),
    data = breast_cancer)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         -2.6744     0.4191  -6.381 1.76e-10 ***
as.numeric(node.caps)   1.4448     0.3132   4.613 3.98e-06 ***
---
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: 313.33  on 275  degrees of freedom
AIC: 317.33

Number of Fisher Scoring iterations: 4
```

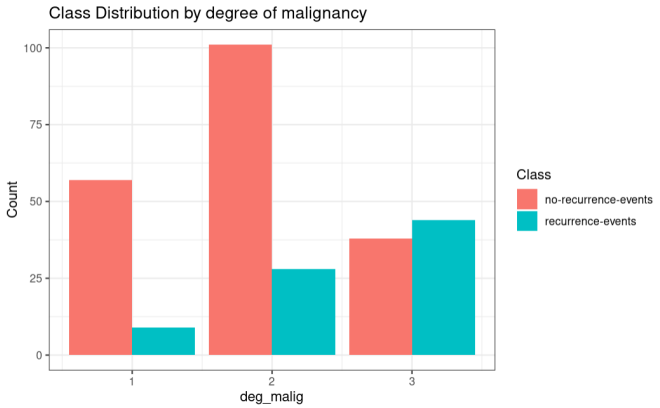
Node encapsulation in breast cancer or Encapsulated papillary carcinoma (EPC) is a rare malignant papillary breast tumor. So, this binary variable (Node_caps) indicates whether the individual has or not node encapsulation.

In our plot, we can see there are more individuals that do not have recurrence if they did not have node encapsulation. And in our model, for the predictor ‘node_caps’, we obtain a p-value<0.05. Meaning that node encapsulation is statistically significant to the outcome of recurrence. So, the logit function is:

$$logit(node.caps) = -2.6744 + 1.448 \times node.caps$$

The coefficient has a positive sign so, as the variable of ‘node_caps’ increases the likelihood of the outcome of recurrence also increases. Also, the odds ratio is 4.2408 this means that for every one-unit increase in the ‘node_caps’ , the odds of the recurrence will increase by 4.2408.

DEGREE OF MALIGNANCY:



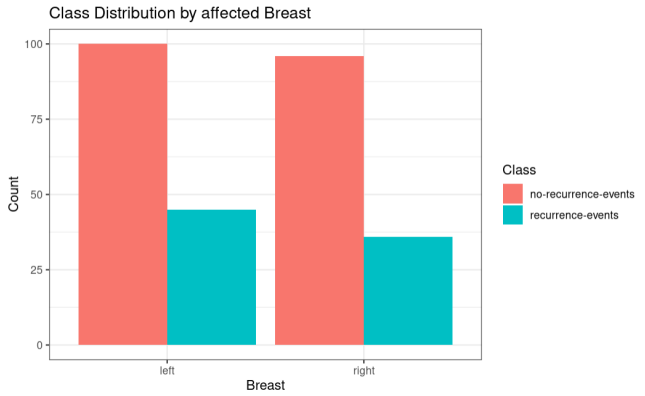

```
Call:
glm(formula = Class ~ deg.malig, family = binomial(link = "logit"),
    data = breast_cancer)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.3009     0.5036  -6.555 5.57e-11 ***
deg.malig     1.1108     0.2116   5.248 1.54e-07 ***
---
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: 302.85  on 275  degrees of freedom
AIC: 306.85

Number of Fisher Scoring iterations: 4
```

We know that the degree of malignancy of a tumor is one of the higher risks of recurrence in breast cancer patients. As we can see in the plot, the higher the degree of malignancy, the more cases of recurrence observed in the dataset. However, it is not the same for no-recurrence, the most frequent is at 2, then 1, and then 3. Additionally, the model for malignancy shows a very small p-value (<0.05) which indicates a statistically significant to the likelihood of recurrence.

The logit function is: $$logit(deg.malig) = -3.3009 + 1.1108 \times deg.malig$$

Once again, we see that the coefficient is positive, showing a positive relationship between the two variables, as the degree of malignancy of the tumor increases as does the likelihood of recurrence. Also, the odds ratio for these two is 3.036635 meaning that for every one-unit increase in the 'deg.malig' , the odds of the recurrence will increase by 3.036635.

AFFECTED BREAST:

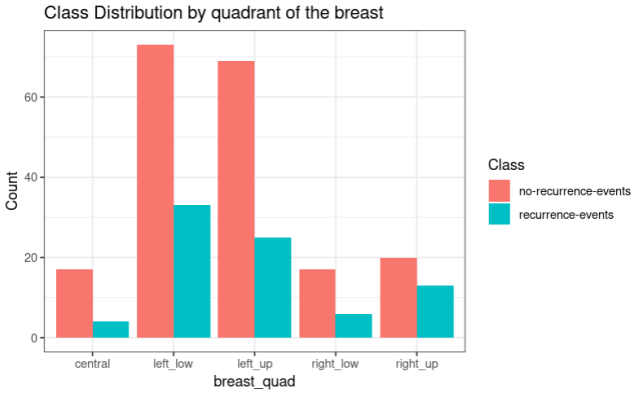



```
Call:
glm(formula = Class ~ as.numeric(breast), family = binomial(link = "logit"),
    data = breast_cancer)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -0.6162     0.4088  -1.507    0.132
as.numeric(breast)  -0.1823     0.2654  -0.687    0.492

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 334.78  on 276  degrees of freedom
Residual deviance: 334.31  on 275  degrees of freedom
AIC: 338.31

Number of Fisher Scoring iterations: 4
```

By observing the plot comparing the affected breast (left or right), we can see that there isn't any difference, in fact the recurrence events are very similar in both cases. In this logistic regression study evaluating breast cancer recurrence, the affected breast (left or right) exhibited non-significant influence on the likelihood of recurrence ($p > 0.05$). Both the intercept (-0.6162, $p = 0.132$) and the breast coefficient (-0.1823, $p = 0.492$) lacked statistical significance.

The distinction between left and right breasts does not significantly predict recurrence events. Therefore we can accept the null hypothesis and state that there is not enough evidence to suggest that 'breast' significantly predicts the Class variable in the model.

QUADRANT OF THE BREAST:




```
Call:
glm(formula = Class ~ as.numeric(breast.quad), family = binomial(link = "logit"),
    data = breast_cancer)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)            -1.2170     0.3639  -3.344 0.000825 ***
as.numeric(breast.quad)   0.1184     0.1192   0.993 0.320571
---
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: 333.80  on 275  degrees of freedom
AIC: 337.8

Number of Fisher Scoring iterations: 4
```

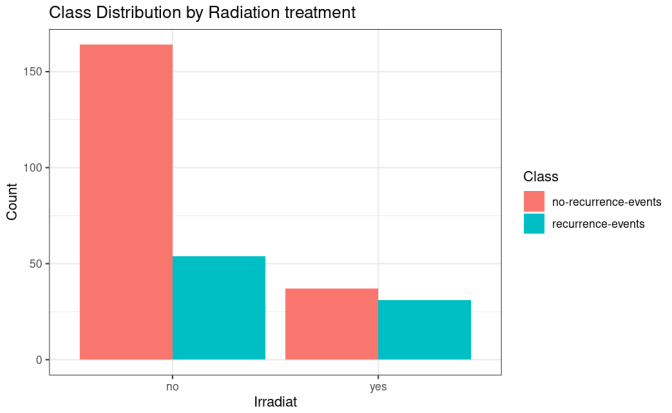
In the case for the breast quadrant, we can see that a tumor in the left low quadrant of the breast has more recurrence than other parts of the breast. This also is the case for the no-recurrence events, the highest of no recurrence is when the tumor was in the left low quadrant. Also, we can see that this variable follows a normal distribution.

Moreover, the model shows that the coefficient of 'breast.quad' is not statistically significant to the likelihood of recurrence as the p-value is bigger than the common significance value of 0.05.

The logit function is: $logit(breast.quad) = -1.2170 + 0.1184 \times breast.quad$

This function with a positive sign in the coefficient shows a small but positive relationship between the two variables and the odds ratio of 1.125681 means that for every one-unit increase in the 'breast.quad' , the odds of the recurrence will increase by 1.125681.

RADIATION TREATMENT:




```
Call:
glm(formula = Class ~ as.numeric(irradiat), family = binomial(link = "logit"),
    data = breast_cancer)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -2.2715     0.4092  -5.552 2.83e-08 ***
as.numeric(irradiat)   1.1035     0.3005   3.672  0.00024 ***
---
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: 321.46  on 275  degrees of freedom
AIC: 325.46

Number of Fisher Scoring iterations: 4
```

Finally, in our plot for the radiation treatment, we can see that there are a lot more individuals with no radiation treatment compared to those who received radiation. We can observe that those who have received radiation seem to have similar probabilities of recurrence or not recurrence. Instead those who haven't received radiation seem to have more probabilities of no recurrence.

This plot has a reason why it is like this. Patients with a less aggressive or early-stage tumor might not necessarily need a radiation treatment resulting in a higher count of "no radiation" group. Indicating that without undergoing radiation, it is enough to surpass cancer. Furthermore, receiving a more aggressive treatment does not mean not having the risk of recurrence due to the aggressive nature of the tumors resulting in individuals with recurrence even though they did the radiation treatment.

Moving on to the model, by looking at the p-values, we can see that the presence of irradiation significantly influences the likelihood of breast cancer recurrence (p-value > 0.05). The model suggests that individuals receiving irradiation have a higher log-odds of experiencing recurrence compared to those without irradiation. The highly significant coefficients and deviance information indicate a strong fit between irradiation and the response variable.

Logit function: $logit(irradiat) = -2.2715 + 1.1035 \times irradiat$

A positive sign in the coefficient indicates a positive relationship between the two variables, as one increases so does the other. And, with an odds ratio of 3.014706, for every one-unit increase in the 'irradiat' , the odds of the recurrence will increase by 3.014706 .

3. Techniques used to analyze the dataset

In our case, our target variable is binary (1 or 2), representing the recurrence or no-recurrence of breast cancer. To analyze this binary outcome, we will employ logistic regression models.
Logistic regression models are well-suited for scenarios where the dependent variable is binary, offering a way to model the probability of an event occurring as a function of predictor variables. Mathematically, the logistic regression model is expressed as:

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Here, p represents the probability of the event (in our case, breast cancer recurrence), and X1,X2,…,Xk are the predictor variables.

To ensure the robustness and appropriateness of our models, we incorporate likelihood ratio tests into our analytical framework.
The likelihood ratio compares the likelihood of the data under the full model against the likelihood under a reduced model, providing a statistical basis for model selection. The test statistic is calculated as:

$$\text{Likelihood Ratio (LR)} = -2 \times (loglikelihood_{reduced} - loglikelihood_{full})$$

Under the null hypothesis that the reduced model is sufficient, this statistic follows a chi-square distribution with degrees of freedom equal to the difference in the number of parameters between the full and reduced models.

By using logistic regression and likelihood ratio tests, we're trying to understand how different factors relate to breast cancer coming back or not. These statistical methods help us pick the best models to study this. We want to make sure our study is done in a careful and insightful way, so we can learn more about what might predict breast cancer recurrence in real-world situations.

4. Results of the analysis

4.1. Generalized linear model

Our dataset contains categorical variables, and an error might be caused by the presence of non-numeric values in a column that is being treated as a numeric predictor in the linear model. We have categorical variables like 'menopause', 'node-caps', 'breast', 'breast-quad', and 'irradiant' that need to be converted into factors.

By looking at the generalized linear model we observe that only two variables are significant (p-value lower than our significance level of 0,05), inv-nodes and deg-malig. We were already expecting to see these results because when we previously plotted individually the variables against class, we observed a very strong relationship between the degree of malignancy and recurrence events. However, in the previous individual analysis (comparing each predictor to Class) we got other

predictors with statistical significance that are not shown in the full model, this can be explained due to some correlation between variables.

```
Call:
glm(formula = Class ~ as.numeric(Age) + as.numeric(Menopause) +
    as.numeric(Tumor_Size) + as.numeric(inv.nodes) + node.caps +
    deg.malig + breast + breast.quad + irradiat, family = binomial(link = "logit"),
    data = breast_cancer)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           -4.56467    1.51439  -3.014 0.002577 **
as.numeric(Age)        0.04149    0.20375   0.204 0.838644
as.numeric(Menopause)  0.20924    0.21063   0.993 0.320517
as.numeric(Tumor_Size) 0.06734    0.06943   0.970 0.332107
as.numeric(inv.nodes)  0.19264    0.08151   2.363 0.018109 *
node.caps              0.23254    0.34752   0.669 0.503399
deg.malig              0.76551    0.22067   3.469 0.000522 ***
breast                -0.32919    0.29443  -1.118 0.263533
breast.quad            0.05599    0.13328   0.420 0.674445
irradiat               0.34019    0.33090   1.028 0.303921
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 348.05  on 285  degrees of freedom
Residual deviance: 302.65  on 276  degrees of freedom
AIC: 322.65

Number of Fisher Scoring iterations: 4
```

Now we can improve our model by eliminating thos non-significant predictors and using only those that really have an impact on the recursion of breast cancer.

```
Call:
glm(formula = Class ~ +deg.malig + as.numeric(inv.nodes), family = binomial(link =
"logit"),
    data = breast_cancer)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6121  -0.6969  -0.6969   0.9751   2.1090

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -3.17017    0.48647  -6.517 7.19e-11 ***
deg.malig             0.81822    0.20703   3.952 7.75e-05 ***
as.numeric(inv.nodes)  0.24238    0.06778   3.576 0.000349 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 348.05  on 285  degrees of freedom
Residual deviance: 308.40  on 283  degrees of freedom
AIC: 314.4

Number of Fisher Scoring iterations: 4
```

If we compare the AIC values we observe that the reduced model has a lower value meaning that it has improved. We can also perform a likelihood ratio test to see if there is improvement:

```
Likelihood ratio test

Model 1: Class ~ +deg.malig + as.numeric(inv.nodes)
Model 2: Class ~ as.numeric(Age) + as.numeric(Menopause) + as.numeric(Tumor_Size) +
    as.numeric(inv.nodes) + as.numeric(node.caps) + deg.malig +
    breast + as.numeric(breast.quad) + irradiat
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   3 -154.20
2  10 -151.33  7 5.7524     0.5689
```

- **Null Hypothesis H_0:** The additional predictors in the full model have no effect, and the simpler model is sufficient.
- **Alternative Hypothesis H_1:** The additional predictors in the full model have a significant effect, and the more complex model is warranted.

The null hypothesis is that the reduced model (Model 1) is sufficient, and the additional parameters in the full model (Model 2) do not significantly improve the fit. The p-value of 0.5689 suggests that we do not have enough evidence to reject the null hypothesis. Therefore, based on this test, we would conclude that the simpler model is adequate for explaining the data.

Now we can generate predictions using the reduced model and then create a confusion matrix to evaluate the performance of the model. This analysis is particularly useful for evaluating the classification performance of the reduced model, comparing its predictions to the actual observed values in the dataset. The confusion matrix helps assess the accuracy and reliability of the model's predictions in distinguishing between different classes.

	observed	
predicted	1	2
0	140	28
1	61	57

The confusion matrix shows the classification results of the reduced model for predicting breast cancer recurrence. Among 228 instances, the model correctly identified 140 cases of non-recurrence (true negatives) and 57 cases of recurrence (true positives). However, there were 28 instances misclassified as non-recurrence (false positives) and 61 instances misclassified as recurrence (false negatives).

To calculate the overall accuracy of the model using the confusion matrix, you can use the following formula:

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Instances} = \frac{140 + 57}{140 + 28 + 61 + 57} = 0,6895$$

So, the overall accuracy of the model is approximately 68.95%.

5. Conclusions and discussions

Our thorough analysis sheds light on the factors influencing breast cancer recurrence and after reducing our full model with all the predictors, we get that the degree of malignancy in a tumor and the involvement of nodes, these are the most influential predictors. Therefore, the degree of malignancy significantly and the involvement of nodes influence whether the cancer might come back or not (recurrence).

In conclusion, these findings emphasize what we had expected at the beginning and the importance of paying close attention to how aggressive tumors are and the involvement of nodes to tailor treatment approaches accordingly to each case. It's a crucial insight for doctors and researchers, suggesting that understanding and managing tumor aggressiveness and node involvement is key in predicting and addressing the risk of breast cancer recurrence.

6. Bibliography

Zwitter,Matjaz and Soklic,Milan. (1988). Breast Cancer. UCI Machine Learning Repository. https://doi.org/10.24432/C51P4M.

Breast cancer. (2023, July 12). World Health Organization (WHO). https://www.who.int/news-room/fact-sheets/detail/breast-cancer

Overview - - - Breast cancer in women. (n.d). NHS. https://www.nhs.uk/conditions/breast-cancer/

Recurrent breast cancer - Symptoms and causes. (2022, July 2). Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/recurrent-breast-cancer/symptoms-causes/syc-20377135

Factors associated with breast cancer recurrences or mortality and dynamic prediction of death using history of cancer recurrences: the French E3N cohort. (2018, February 9). NCBI. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5807734/

7. Appendix

Libraries:

```
library(ggplot2)
library(lmtest)
```

Load data and study it:

```
breast_cancer <- read.csv("breast+cancer/breast-cancer.data")
nrow(breast_cancer)

breast_cancer$Age <- as.factor(breast_cancer$Age)
breast_cancer$Class <- as.factor(breast_cancer$Class)
breast_cancer$Menopause <- as.factor(breast_cancer$Menopause)
breast_cancer$Tumor_Size <- as.factor(breast_cancer$Tumor_Size)
breast_cancer$inv.nodes <- as.factor(breast_cancer$inv.nodes)
breast_cancer$node.caps <- as.factor(breast_cancer$node.caps)
breast_cancer$breast <- as.factor(breast_cancer$breast)
breast_cancer$breast.quad <- as.factor(breast_cancer$breast.quad)
breast_cancer$irradiat <- as.factor(breast_cancer$irradiat)

summary(breast_cancer$Class)
```

Remove missing values:

```
breast_cancer[breast_cancer == '?'] <- NA
breast_cancer <- na.omit(breast_cancer)
```

Plots class against variables:

```
ggplot(breast_cancer, aes(x = Class, fill = Class)) +
  geom_bar() +
  labs(title = "Class Distribution") +
  theme_bw()

ggplot(breast_cancer, aes(x = Age, fill = Class)) +
  geom_bar(position = "dodge") +
  labs(title = "Class Distribution by Age Group", x = "Age Group", y = "Count", fill = "Class") +
  theme_bw()

ggplot(breast_cancer, aes(x = Menopause, fill = Class)) +
  geom_bar(position = "dodge") +
  labs(title = "Class Distribution by Menopause Group", x = "Menopause Group", y = "Count", fill =
"Class") +
  theme_bw()

ggplot(breast_cancer, aes(x = Tumor_Size, fill = Class)) +
  geom_bar(position = "dodge") +
  labs(title = "Class Distribution by Tumor Size", x = "Tumor Size", y = "Count", fill = "Class") +
  theme_bw()

ggplot(breast_cancer, aes(x = inv.nodes, fill = Class)) +
  geom_bar(position = "dodge") +
  labs(title = "Class Distribution by involved lymph nodes", x = "inv_nodes", y = "Count", fill =
"Class")+
  theme_bw()

ggplot(breast_cancer, aes(x = node.caps, fill = Class)) +
  geom_bar(position = "dodge") +
  labs(title = "Class Distribution by node capsulation", x = "Node_caps", y = "Count", fill = "Class") +
  theme_bw()


ggplot(breast_cancer, aes(x = deg.malig , fill = Class)) +
  geom_bar(position = "dodge") +
  labs(title = "Class Distribution by degree of malignancy", x = "deg_malig", y = "Count", fill =
"Class") +  theme_bw()

ggplot(breast_cancer, aes(x = breast, fill = Class)) +
  geom_bar(position = "dodge") +
  labs(title = "Class Distribution by affected Breast", x = "Breast", y = "Count", fill = "Class") +
  theme_bw()

ggplot(breast_cancer, aes(x = breast.quad, fill = Class)) +
  geom_bar(position = "dodge") +
  labs(title = "Class Distribution by quadrant of the breast", x = "breast_quad", y = "Count", fill =
"Class") +
  theme_bw()


ggplot(breast_cancer, aes(x = irradiat, fill = Class)) +
```

```
geom_bar(position = "dodge") +
labs(title = "Class Distribution by Radiation treatment", x = "Irradiat", y = "Count", fill = "Class") +
theme_bw()
```

Linear models

```
# AGE
        model_age <- glm(Class~as.numeric(Age), data= breast_cancer, family = binomial(link =
        "logit"))
        summary(model2)
        OR_age <- exp(coefficients(model_age)[2])
        OR_age

# MENOPAUSE
        model_meno <-glm(Class~as.numeric(Menopause), data= breast_cancer, family =
        binomial(link = "logit"))
        summary(model_meno)
        OR_meno <- exp(coefficients(model_meno)[2])
        OR_meno

# TUMOR SIZE
        model_tumor <-glm(Class~as.numeric(Tumor_Size), data= breast_cancer, family =
        binomial(link = "logit"))
        summary(model_tumor)
        OR_tumor <- exp(coefficients(model_tumor)[2])
        OR_tumor


# NODE INVOLVEMENT
        model_nodes <-glm(Class~as.numeric(inv.nodes), data= breast_cancer, family = binomial(link
        = "logit"))
        summary(model_nodes)
        OR_inv <- exp(coefficients(model_nodes)[2])
        OR_inv

# NODE ENCAPSULATION
        model_node_caps <-glm(Class~as.numeric(node.caps), data= breast_cancer, family =
        binomial(link = "logit"))
        summary(model_node_caps)
        OR_node_caps <- exp(coefficients(model_node_caps)[2])
        OR_node_caps

# DEGREE OF MALIGNANCY
        model_malig <-glm(Class~deg.malig, data= breast_cancer, family = binomial(link = "logit"))
        summary(model_malig)
        OR_malig <- exp(coefficients(model_malig)[2])
```

```
OR_malig
# BREAST
model_breast <- glm(Class~as.numeric(breast), data= breast_cancer, family = binomial(link =
"logit"))
summary(model_breast)
OR_breast <- exp(coefficients(model_breast)[2])
OR_breast

# BREAST QUADRANT
model_breast_q <- glm(Class~as.numeric(breast.quad), data= breast_cancer, family =
binomial(link = "logit"))
summary(model_breast_q)
OR_breast_q <- exp(coefficients(model_breast_q)[2])
OR_breast_q

# IRRADIATION
model_irradiat <- glm(Class~as.numeric(irradiat), data= breast_cancer, family = binomial(link
= "logit"))
summary(model_irradiat)
OR_irr <- exp(coefficients(model_irradiat)[2])
OR_irr
```

Models:

```
# FULL MODEL
full_model <- glm(Class ~ as.numeric(Age) + as.numeric(Menopause) +
as.numeric(Tumor_Size) + as.numeric(inv.nodes) + as.numeric(node.caps) + deg.malig +
breast + as.numeric(breast.quad) + irradiat, data = breast_cancer, family = binomial(link =
"logit"))
summary(full_model)

# REDUCED MODEL
reduced_model <- glm(Class ~ + deg.malig + as.numeric(inv.nodes), data = breast_cancer,
family = binomial(link = "logit"))
summary(reduced_model)

# LIKELIHOOD RATIO TEST
lr_test <- lrtest(reduced_model, full_model)
print(lr_test)

# CONFUSION MATRIX
pred <- predict(reduced_model, type = "response")
predicted <-ifelse(pred > mean(pred), 1,0)
table_q <- table(predicted, observed = as.numeric(breast_cancer$Class))
table_q
```