Data ScienceTech Institute

Breast cancer dataset analysis using semi parametric cox regression model

Prepared by

Djihene Beladjine
Josselin Cachet
Deena Zamzam

2023

TABLE OF CONTENTS

1) Introduction

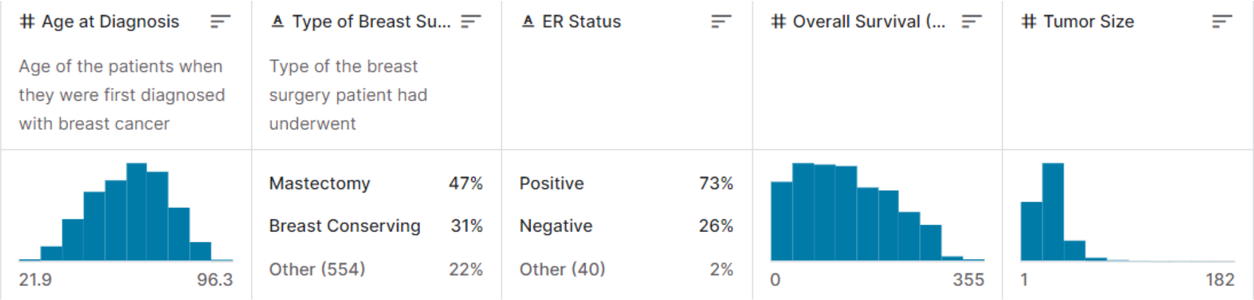
In this Survival Analysis project, we performed a survival analysis in R using a semi-parametric Cox regression model on a breast Cancer dataset which can be found at the following address : https://www.kaggle.com/datasets/gunesevitan/breast-cancer-metabric/data

Our database contains the clinical profiles of 2,509 breast cancer patients, with 34 columns including sex, ER status, Chemotherapy, overall survival status, Cancer type, tumor size and so on.

The project was hosted on Github.

2) Data Loading, Exploration and Cleaning

The main requirement of a dataset in survival analysis is that it needs to be time-to-event data. Our dataset has continuous variable as survival Months, and dichotomous variable as "Living" or "deceased". Here is an overview of the dataset.



# Age at Diagnosis	A Type of Breast Su...	A ER Status	# Overall Survival (...	# Tumor Size
Age of the patients when they were first diagnosed with breast cancer	Type of the breast surgery patient had underwent			
	Mastectomy 47%	Positive 73%		
	Breast Conserving 31%	Negative 26%		
21.9 – 96.3	Other (554) 22%	Other (40) 2%	0 – 355	1 – 182

The data is manipulated and assessed in R and RStudio. The functions used to do so are derived from existing packages: survival, broom, ggplot2, vctrs, dplyr, survmier, tidyverse.

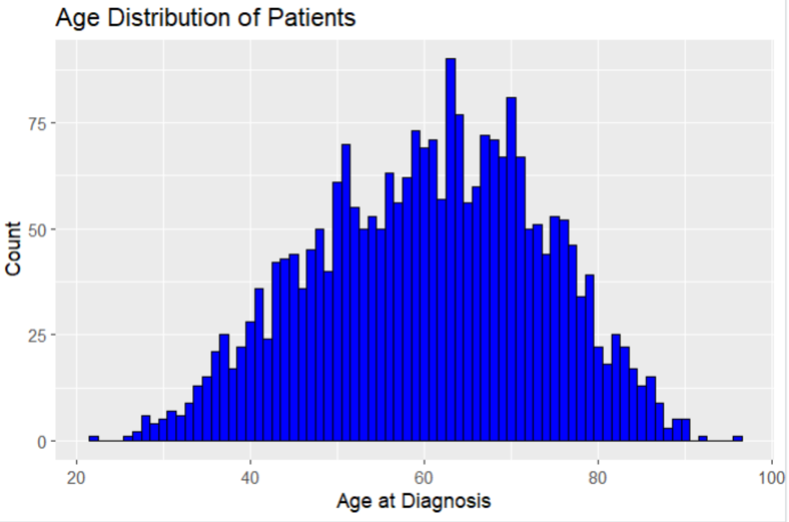
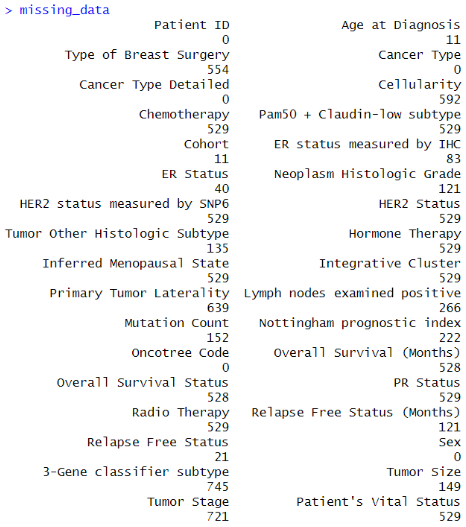


Figure: Age Distribution

Handling the missing values:

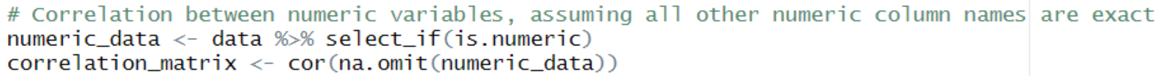
For handling the missing values, we used "sapply" function



```
> missing_data
                   Patient ID              Age at Diagnosis
                            0                            11
       Type of Breast Surgery                   Cancer Type
                          554                             0
         Cancer Type Detailed                   Cellularity
                            0                           592
                 Chemotherapy    Pam50 + Claudin-low subtype
                          529                           529
                       Cohort        ER status measured by IHC
                           11                            83
                    ER Status      Neoplasm Histologic Grade
                           40                           121
    HER2 status measured by SNP6                HER2 Status
                          529                           529
  Tumor Other Histologic Subtype              Hormone Therapy
                          135                           529
     Inferred Menopausal State          Integrative Cluster
                          529                           529
      Primary Tumor Laterality  Lymph nodes examined positive
                          639                           266
               Mutation Count     Nottingham prognostic index
                          152                           222
                Oncotree Code       Overall Survival (Months)
                            0                           528
       Overall Survival Status                     PR Status
                          528                           529
                 Radio Therapy     Relapse Free Status (Months)
                          529                           121
           Relapse Free Status                           Sex
                           21                             0
     3-Gene classifier subtype                    Tumor Size
                          745                           149
                   Tumor Stage          Patient's Vital Status
                          721                           529
```

Figure: Missing values of each column

Then, we used a correlation function to compensate the missing values.



```
# Correlation between numeric variables, assuming all other numeric column names are exact
numeric_data <- data %>% select_if(is.numeric)
correlation_matrix <- cor(na.omit(numeric_data))
```

3) Survival Data Setup

In this step, we converted `Overall Survival Status` to numeric, and assigned to 'died_of_cancer'. In this study, we specified the factor `Overall Survival (Months)` as the survival time variable and the variable 'died_of_cancer' for the binary event indicator variable.

The plot below represents the varying covariates over time:

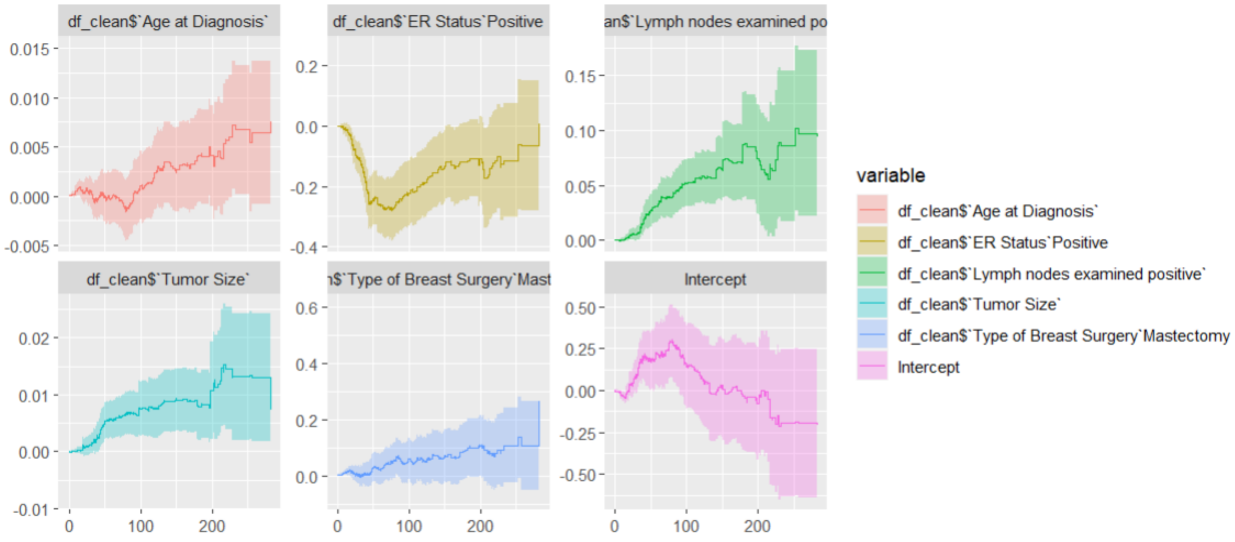


Figure 1 Covariates overtime

4) Covariates Selection

Age at Diagnosis: Age is an important factor in cancer survival analysis.
Type of Breast Surgery: this factor might impact survival.
Tumor Size: this factor provides information about the extent and growth of the tumor.
ER status (Estrogen Receptor Status): ER Status are proteins found on the surface of the breast cells that can bind to estrogen hormones. This factor can influence both prognosis and treatment decisions.
Lymph Nodes Examined Positive: The presence of lymph node indicates the spread of the cancer.

5) Fit Cox Regression Model

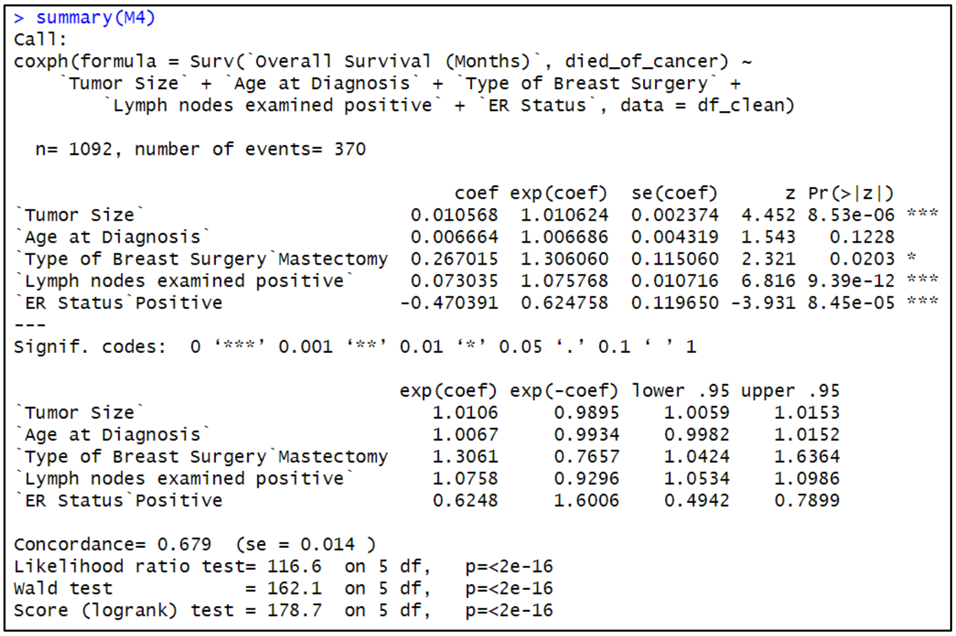
The primary purpose of survival analysis is to study the relationship between variable X and survival function. The Semi-parametric Cox regression is a statistical method used in survival analysis to investigate the relationship between risk factors and survival time. It is a variant of the Cox regression model that allows for the inclusion of both parametric and non-parametric components.

In this step, we are using the coxph() function from the survival package to fit the Cox regression model. We used 4 models including the baseline model for M0. Models M1, M2 and M3 are univariate models:

- For M1 we used the covariate 'Age at Diagnosis',
- For M2 we used the covariate 'Type of Breast Surgery',
- For M3 we used the covariate 'Tumor Size',

For the model M4, we used a multivariate model with the covariates: `Tumor Size`, `Age at Diagnosis`, `Type of Breast Surgery`, `Lymph nodes examined positive` and `ER Status`.

Here is below the summary of the cox model for model M4:



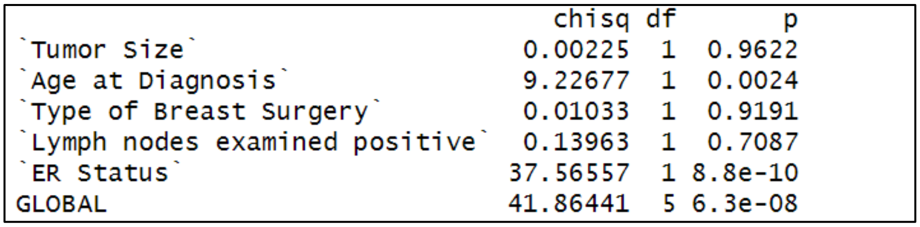
```
> summary(M4)
Call:
coxph(formula = Surv(`Overall Survival (Months)`, died_of_cancer) ~
    `Tumor Size` + `Age at Diagnosis` + `Type of Breast Surgery` +
        `Lymph nodes examined positive` + `ER Status`, data = df_clean)

  n= 1092, number of events= 370

                                        coef exp(coef)  se(coef)      z Pr(>|z|)
`Tumor Size`                        0.010568  1.010624  0.002374  4.452 8.53e-06 ***
`Age at Diagnosis`                  0.006664  1.006686  0.004319  1.543   0.1228
`Type of Breast Surgery`Mastectomy  0.267015  1.306060  0.115060  2.321   0.0203 *
`Lymph nodes examined positive`     0.073035  1.075768  0.010716  6.816 9.39e-12 ***
`ER Status`Positive                -0.470391  0.624758  0.119650 -3.931 8.45e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                                   exp(coef) exp(-coef) lower .95 upper .95
`Tumor Size`                          1.0106     0.9895    1.0059    1.0153
`Age at Diagnosis`                    1.0067     0.9934    0.9982    1.0152
`Type of Breast Surgery`Mastectomy    1.3061     0.7657    1.0424    1.6364
`Lymph nodes examined positive`       1.0758     0.9296    1.0534    1.0986
`ER Status`Positive                   0.6248     1.6006    0.4942    0.7899

Concordance= 0.679  (se = 0.014 )
Likelihood ratio test= 116.6  on 5 df,   p=<2e-16
Wald test            = 162.1  on 5 df,   p=<2e-16
Score (logrank) test = 178.7  on 5 df,   p=<2e-16
```

6) Interpretation of the results

Once the mode is fitted, we can extract and interpret the results:

- The coefficient Estimates (β): The column 'coef', these values represent the log hazard ratio. For the factors 'Age at Diagnosis', 'Tumor Size', 'Type of Breast Surgery' for mastectomy, 'Lymph nodes examined positive', their coefficients are positive which mean an increased hazard. And for the factor 'ER Status' positive, its coefficient is negative which applies a decreased hazard.
- Hazard Ratios (HR): the column 'exp coef', the exponential of the coefficient which is the Hazard ratio. When a Hazard ration is greater than 1, it indicates an increased risk (a higher hazard), in our case the factors who have a positive coefficient Estimates, have also an increased risk. Meanwhile the Hazard ratio of 'ER status' factor is less than 1 which indicates a decreased risk (a protective effect).
- se(coef): The standard error of the coefficient. It measures the variability or uncertainty in the estimated coefficient.
- Z: The z-value is the coefficient divided by its standard error. It is used to test the null hypothesis that the coefficient equals 0.
- Pr(>|z|): The p-value for the z-test. It indicates whether the predictor is statistically significant. A common threshold for significance is 0.05. Here, the p-value is just above 0.05, indicated with a dot (.), which means it is suggestive but not conventionally statistically significant.
- A concordance of 0.679 suggests that the Cox regression model has a strong ability to discriminate between individuals with different survival outcomes. It indicates that the model is reasonably effective at identifying individuals who are more likely to die from cancer within such a dataset.

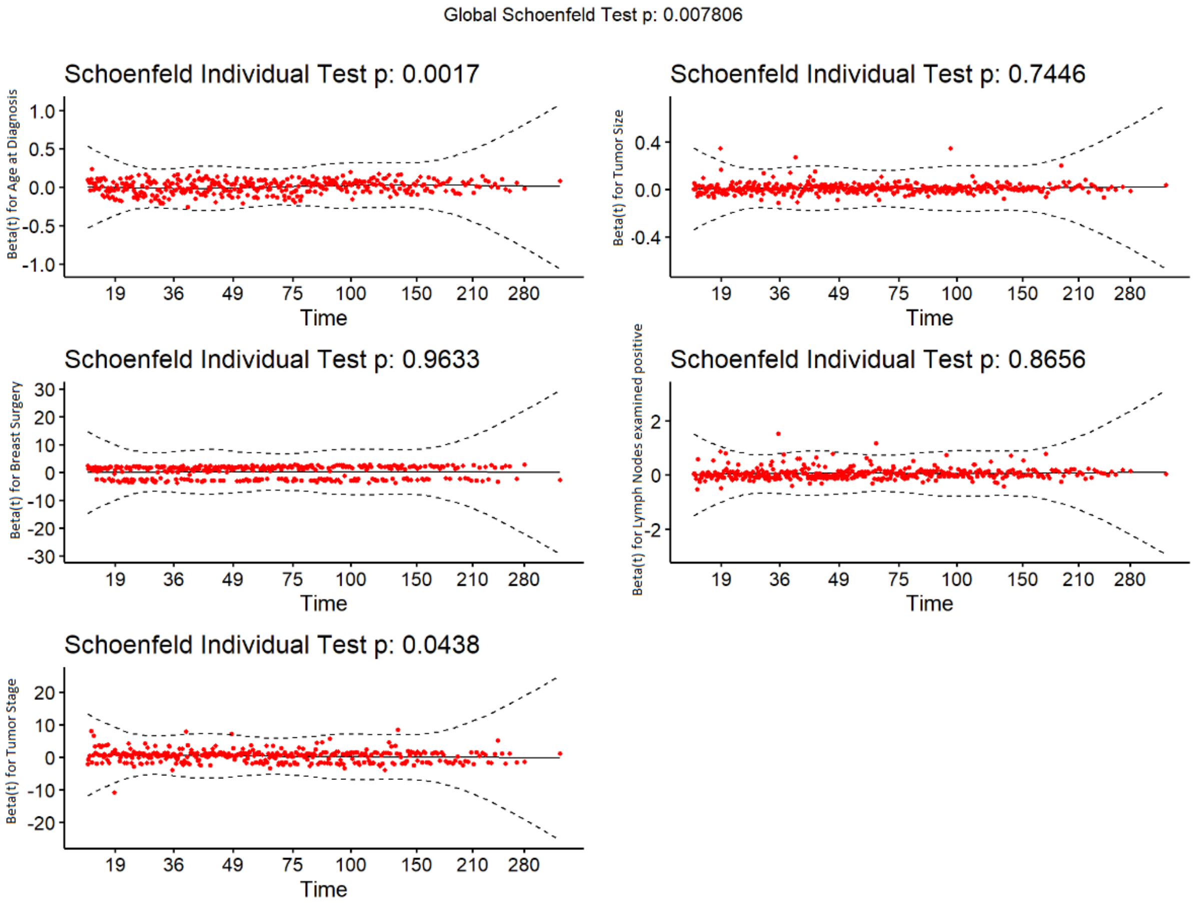
A. PROPORTIONAL HAZARDS ASSUMPTION:



```
                                   chisq df       p
`Tumor Size`                     0.00225  1  0.9622
`Age at Diagnosis`               9.22677  1  0.0024
`Type of Breast Surgery`         0.01033  1  0.9191
`Lymph nodes examined positive`  0.13963  1  0.7087
`ER Status`                     37.56557  1 8.8e-10
GLOBAL                          41.86441  5 6.3e-08
```

The variables with p-values less than the significance level (typically 0.05) can be considered statistically significant. In our case, the significant variables are: « Age at Diagnosis » and « ER Status».

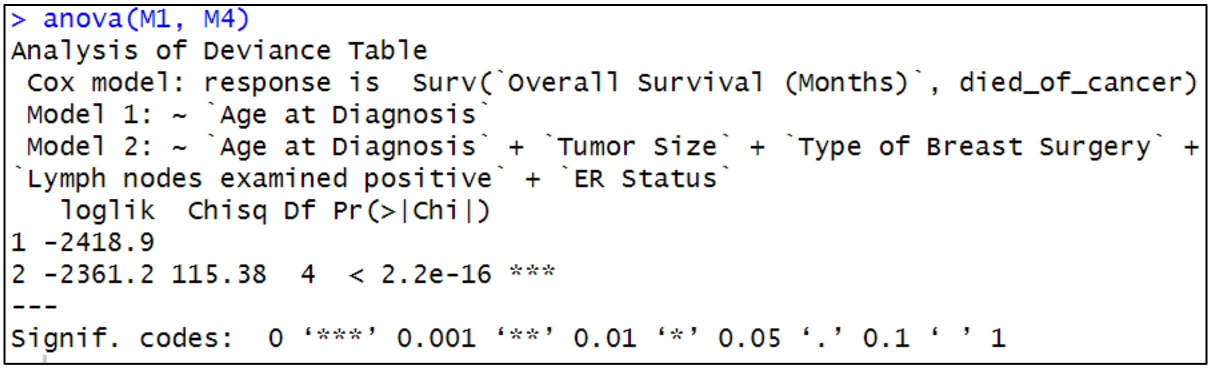
Here is below a visualizing of the scaled Schoenfeld residuals against the transformed time for each covariate:



7) Diagnosis for Model M4

LRT (Likelihood Ratio Test):

It is a test used to compare the fit of two nested models, where one is a reduced version of the other. In the context of Cox regression models, we used the ANOVA () function, as well as models M1 and M4 for the statistical test.

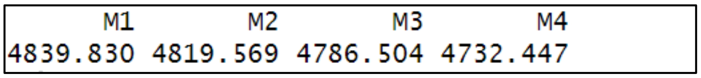


```
> anova(M1, M4)
Analysis of Deviance Table
 Cox model: response is  Surv(`Overall Survival (Months)`, died_of_cancer)
 Model 1: ~ `Age at Diagnosis`
 Model 2: ~ `Age at Diagnosis` + `Tumor Size` + `Type of Breast Surgery` + 
`Lymph nodes examined positive` + `ER Status`
   loglik  Chisq Df Pr(>|Chi|)    
1 -2418.9                         
2 -2361.2 115.38  4  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

These results show us that the Model 2 is significantly better than the Model 1, as indicated by the p-value of model 2 being less than the significance level (typically 0.05). It means that the additional variables included in Model 2 improve the fit of the model.

AIC (Akaike Information Criterion):

It is a measure of the relative quality of statistical models for a given set of data. For our code, we used a list of our Cox regression models.



```
      M1       M2       M3       M4 
4839.830 4819.569 4786.504 4732.447 
```

We noticed that the model M4 has the lowest AIC. In survival analysis, a lower AIC indicates a better-fitting model.

Visualizing the estimated distribution of survival times: Plotting the baseline survival function using the model M4.

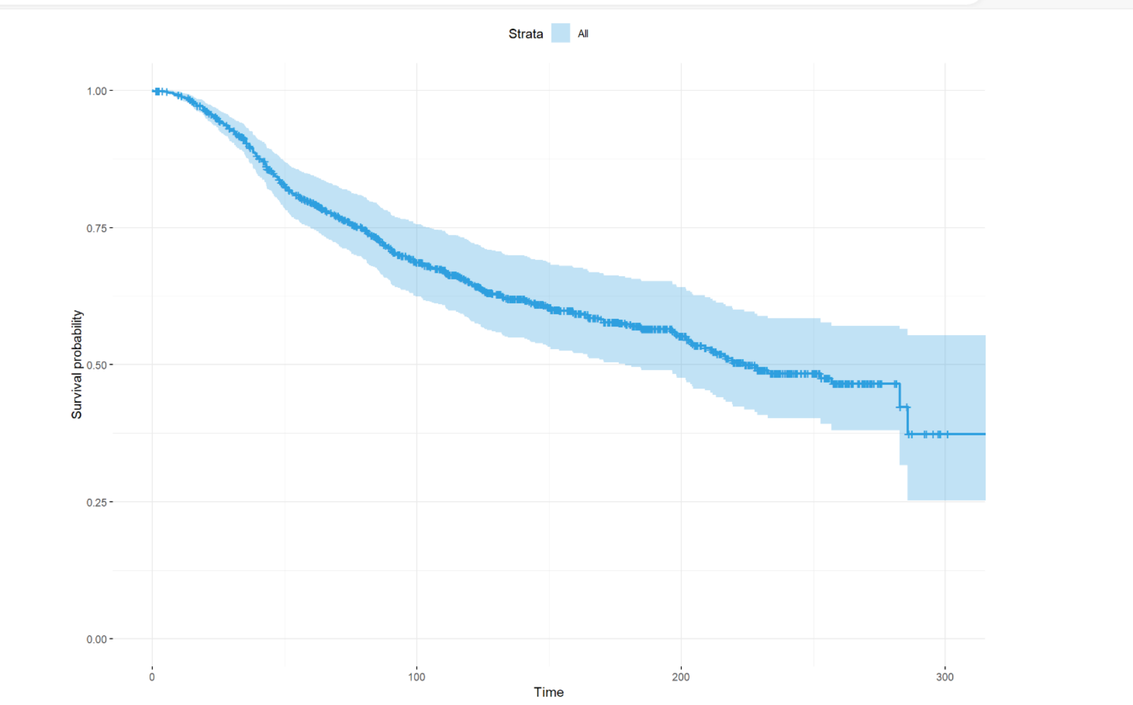


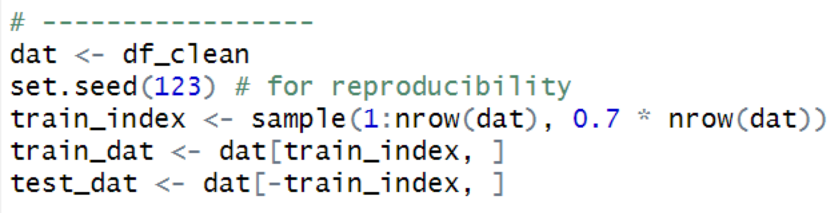
Figure: Cox Proportional Hazards model survival curve.

8) Training, Predicting and Evaluating a model

The first step. We need to process the new dataset and splitting the training set and the testing set:

We assigned the dataset « df_clean » to a new variable « dat ». it is for allowing us to work with the dataset using the « dat » variable instead of the original « df_clean » variable.

The code below is splitting the original dataset (df_clean) into a training set (train_dat) and a testing set (test_dat) for doing our statistical modeling. The training set will be used to build and train a model, while the testing set will be used to evaluate the model with unseen data.



```
# -----------------
dat <- df_clean
set.seed(123) # for reproducibility
train_index <- sample(1:nrow(dat), 0.7 * nrow(dat))
train_dat <- dat[train_index, ]
test_dat <- dat[-train_index, ]
```

A. TRAIN THE NEW MODEL ON THE TRAINING DATASET:

The second step. The training will be done with the Cox regression model. We choose the two covariates: `Lymph nodes examined positive` and `Age at Diagnosis`. The survival time variable is the factor `Overall Survival (Months)`, and the event indicator variable is the factor (died_of_cancer), from the original dataset.

B. MAKING PREDICTION ON THE TRAINING DATASET:

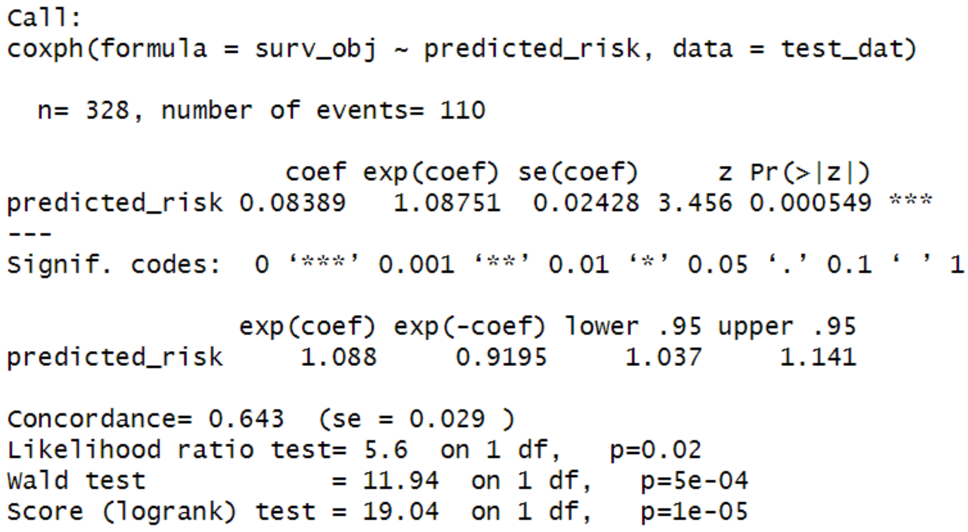
The third step. The testing dataset will have a new column called (predicted_risk), where we will place the result of the prediction, using (predict) function having parameters of the trained model (M4_train), the data used here (test_dat), and precising the type « risk »).

C. EVALUATE THE MODEL:

The final step. Creating a survival object, with the survival time variable is the factor `Overall Survival (Months)`, and the event indicator variable is the factor (died_of_cancer), from the testing dataset (test_dat).

Then, using this survival object, the predicted covariate (predicted_risk), and the testing dataset (test_dat) in the Cox regression function (coxph). The result will be assigned to a model called (cox_fit).

D. MODEL DIAGNOSTICS:



```
Call:
coxph(formula = surv_obj ~ predicted_risk, data = test_dat)

  n= 328, number of events= 110 

                  coef exp(coef) se(coef)     z Pr(>|z|)    
predicted_risk 0.08389   1.08751  0.02428 3.456 0.000549 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

               exp(coef) exp(-coef) lower .95 upper .95
predicted_risk     1.088     0.9195     1.037     1.141

Concordance= 0.643  (se = 0.029 )
Likelihood ratio test= 5.6  on 1 df,   p=0.02
Wald test            = 11.94  on 1 df,   p=5e-04
Score (logrank) test = 19.04  on 1 df,   p=1e-05
```

9) Interpretation

- coef (0.08389): This is the estimated coefficient for predicted_risk. It indicates the log hazard ratio, which is the expected change in the log hazard for a one-unit increase in predicted_risk.
- exp(coef) (1.08751): This is the hazard ratio (HR). HR greater than 1 suggests a higher hazard (or risk) as the predicted_risk increases. Specifically, for each one-unit increase in predicted_risk, the hazard of the event occurring is expected to increase by 9%.
- se(coef) (0.02428): The standard error of the coefficient. It measures the variability or uncertainty in the estimated coefficient.
- z (3.456): The z-value is the coefficient divided by its standard error. It is used to test the null hypothesis that the coefficient equals 0.
- Pr(>|z|) (0.000549): The p-value for the z-test. It indicates whether the predictor is statistically significant. A common threshold for significance is 0.05. Here, the p-value is just above 0.05, indicated with a dot (.), which means it is suggestive but not conventionally statistically significant.

The confidence interval for the hazard ratio:

- Lower .95 (1.037) and Upper .95 (1.141): The 95% confidence interval for the HR does not include 1, which suggests that there may be an effect, but since the p-value is greater than 0.05, we would not typically consider this to be statistically significant.
- Concordance (0.643): The concordance statistic is a measure of the predictive accuracy of the model and ranges from 0.5 (no predictive ability) to 1 (perfect prediction). A value of 0.643 indicates a fair predictive ability.
- Likelihood ratio test (p=0.02): This test compares the goodness of fit of the model against a null model with no predictors. A p-value of 0.1 suggests that the model is significantly better than the null model at the 0.05 level.
- Wald test (p= 5e-04) and Score (logrank) test (p= 1e-05): These are additional tests for the significance of the predictors. Both tests indicate that the observed data is statistically significant.

In conclusion, the predicted_risk variable appears to have a trend towards being a predictor of survival, but this relationship is not statistically significant at the p < 0.05 level. The model has a fair predictive ability as indicated by the concordance statistic.

CONCLUSION

In this study for analyzing the Cox regression model for a breast cancer dataset, we considered two models: a multivariate model (M4) with five clinical factors, and a predictive model centered on the variable ‘predicted_risk’. While the predictive aspect shows promise, a careful consideration of clinical factors in the multivariate aspect shows that it is essential for a comprehensible understanding.

The M4 Model offers a more detailed understanding of the factors influencing breast cancer survival, and its slightly higher concordance suggests improved predictive accuracy compared to the Predictive Model. Further research and validation with medical expertise assistance may be necessary to confirm the significance of certain factors.