Breast Cancer Survival Analysis

Yash Mehta

Abstract

Through this study on the breast cancer dataset from National Cancer Institute which records a patients' history from 1973 to 2014 and have more than 125 attributes. Many of the attributes are categorical variables and hence require a different treatment compared to continuous variable data set. In this research, I have used the application of Multiple correspondence analysis and Hierarchical Clustering to identify the various factors that may affect the survival of a breast cancer patient. To further strengthen the results and compare the results obtained from clustering, I applied the Cox proportional Hazard method to perform survival analysis. The results the various methods were then compared to reach the conclusion.

Introduction

Breast cancer is the most common malignant disease for females and the second most common type of cancer after lung cancer for both sexes. It primarily affects women older than 50 years. Even though the absolute incidence in women aged 20 - 40 years is low, breast cancer constitutes about 24 percent of new cancers in this age group. Hence treatment of breast cancer, including surgery, drugs (hormone therapy and chemotherapy) and radiation, is a main interest of the public health sector. In USA and Canada, breast cancer accounts for 29% of all cancer diagnoses for women. One in 9 women is expected to develop breast cancer during her lifetime and one in 33 is expected to die of breast cancer.

It is hence important to understand several factors that contribute to the survival of a breast cancer patient. A survival rate is a statistical index that summarizes the probable frequency of specific outcomes for a group of patients at a point in time. A survival curve is a summary display of the pattern of survival rates over time. The basic concept is simple. For example, for a certain category of patient, one might ask what proportion is likely to be alive at the end of a specified interval, such as 5 years. The greater the proportion surviving, the lower the risk for this category of patients. Survival analysis, however, is somewhat more complicated than it first might appear. If one were to measure the length of time between diagnosis and death or record the vital status when last observed for every patient in a selected patient group, one might be tempted to describe the survival of the group as the proportion alive at the end of the period under investigation. This simple measure is informative only if all the patients were observed for the same length of time.

The MCA is a dimension reduction technique like factor analysis, but extends factor analysis in two counts; handling of categorical variable, particularly measured in nominal scale and developing perceptual maps of extracted components. It helps us in identifying the crucial factors that contribute to data variability and hence will be used here to recognize the important attributes.

This study is further extended to ATTRIBUTE CLUSTERING where different attributes are factored together to identify the number of clusters that can be formed for the given data set and check inter variable dependency.

The COX PROPORTIONAL HAZARD METHOD further strengthens the study of various data attributes and is and provides important insights for survival analysis. It identifies the crucial factors and provide a fit curve to predict the survival of future cancer cases. It can help in predicting the survival probability of patient and identify the factors that are classified as hazard.

KAPLAN-MEIER METHOD has also been employed. Since the individual patient data is available, these same data can be analyzed using the Kaplan-Meier method. It calculates the proportion surviving to each point that a death occurs, rather than at fixed intervals. The principal difference evident in a survival curve is that the stepwise changes in the cumulative survival rate appear to occur independently of the intervals on the “Years Following Diagnosis” axis. This method provides a more accurate estimate of the survival curve.

Data Used for Analysis

The data set used for this study is fron National Cancer Institute’s Surveillance, Epidemiology, and End Results Program. It records the data of breast cancer patients from around the country of USA from 1973 to 2014. SEER collects and publishes cancer incidence and survival data from population-based cancer registries covering approximately 28% of the population of the United States. SEER coverage includes 26% of African Americans, 41% of Hispanics, 43% of American Indians and Alaska Natives, 54% of Asians, and 71% of Hawaiian/Pacific Islanders. The SEER Program registries routinely collect data on patient demographics, primary tumor site, tumor morphology and stage at diagnosis, first course of treatment, and follow-up for vital status.

The data used is large and hence we used the batch processing techniques to read the data in R studio and used following R- packages

We have a large dataset with more than 100 columns. However, it is important to analyze only the columns (or attributes) that affect the breast cancer patient and can further help in survival analysis. Different columns were manually inspected and a few attributes were shortlisted for initial analysis. The following columns were then shortlisted for end to end analysis:

Sex, Marital.Status, Race.Ethnicity, Primary.Site, Laterality,Histology, Behavior.Code,Grade, HER2.Recode, ER.Status, PR.Status, Breast.Subtype, Vital.Status.recode, Age.at.diagnosis, Survival.months, Total.Number.of.Benign.Tumors, Total.Number.of.In.Situ.malignant.Tumors.

The data is the set of cancer cases from various regions and have ID unique to each region called as “Registry ID”. Here is the count of the number of patients or cases under different regions.

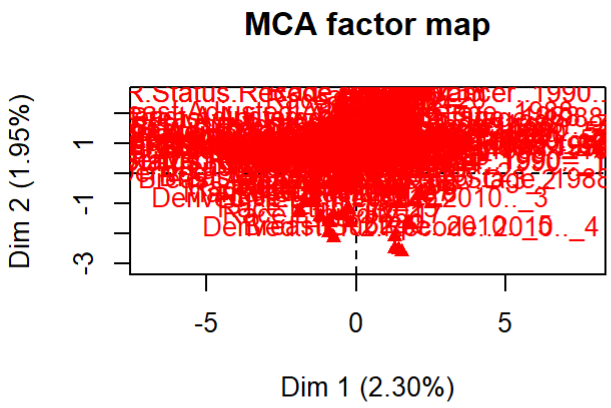
```
##
##   1501   1502   1520   1521   1522   1523   1525   1526   1527
## 131674 127824 132366  34426  95375  42617 123176  41221  69945
```

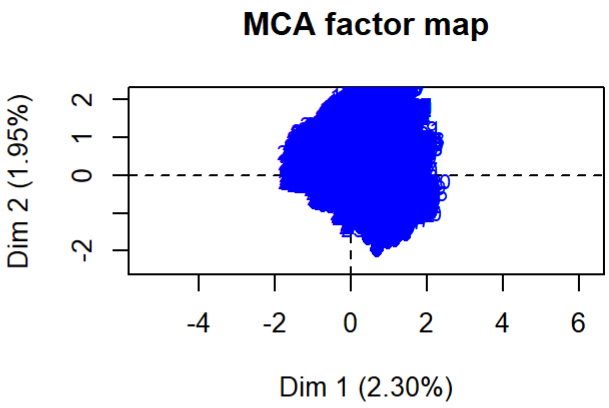
To Identify a smaller dataset and find out the important attributes that explain the variation in breast cancer data we applied Multiple Correspondence Analysis (MCA) on the registry ID with most number of records. MCA is an important dimension reduction technique used for categorical dataset. From the above it can be concluded that registry 1520 has largest number of cases and hence we applied MCA to the same.

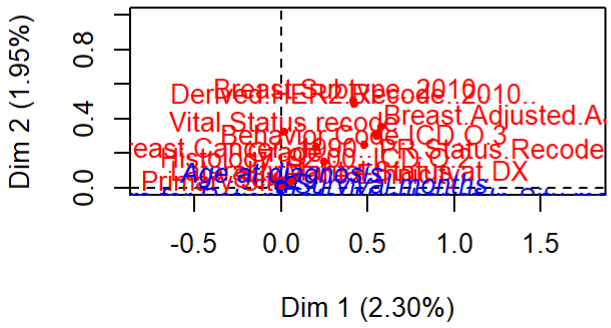
MCA Analysis

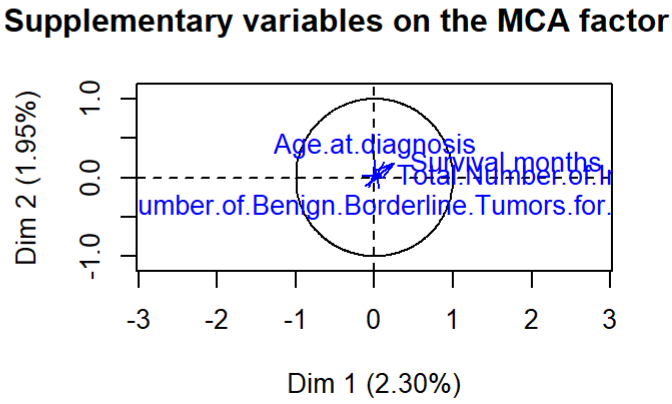
MCA as stated above stands for multiple correspondence analysis and can be implemented in R using FactoMineR package. It is used to convert the categorical dataset in to a set of mutually orthogonal dimensions that explain the data variation. On applying the MCA on the subset of data described above we obtain following

data plots created using first two dimensions. The first two dimensions explain maximum variation in data.







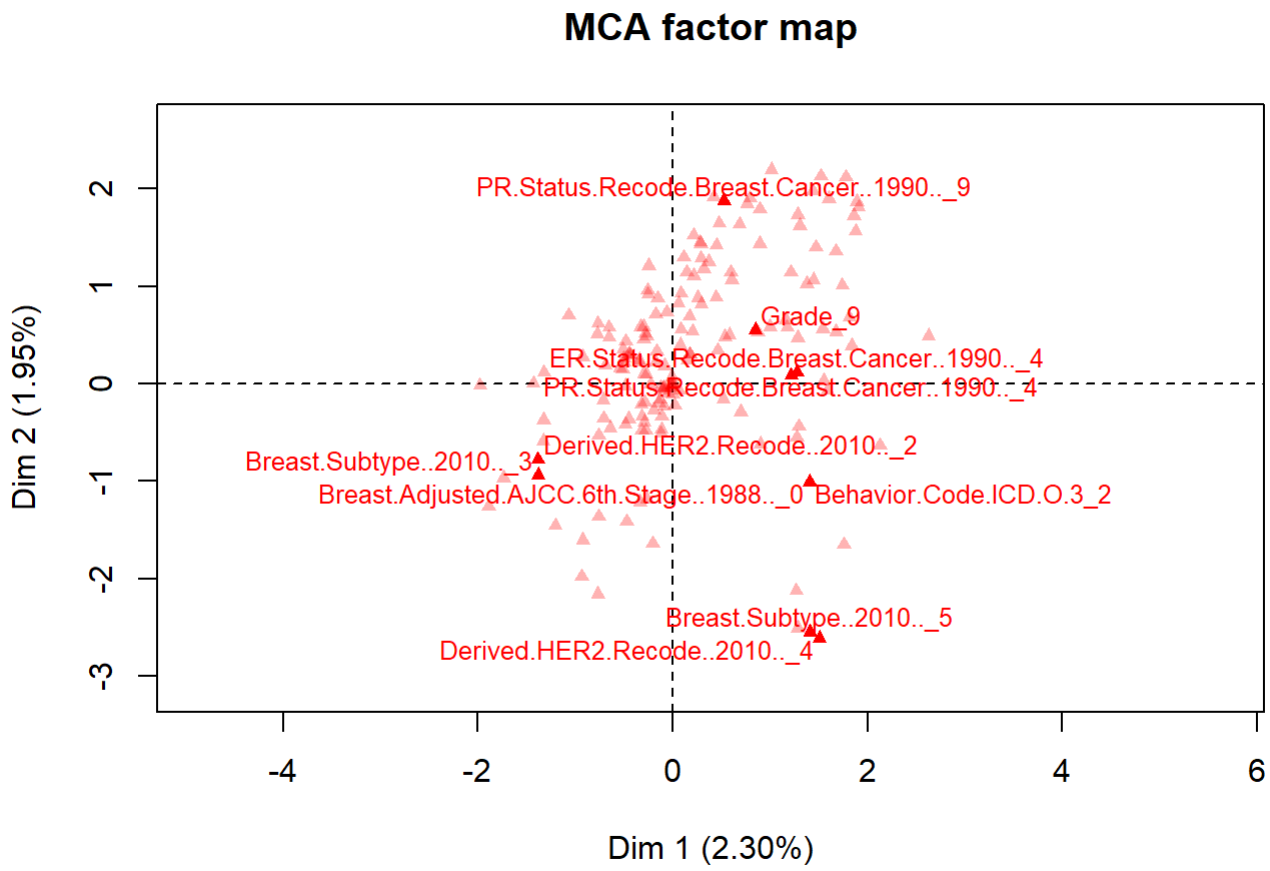


The graphs above plot a lot of data points on top of each other and hence make is difficult for analysis. However using the plot of supplementary variables we can say that the supplementary variables like age at diagnosis , survival months, number of malignant and benign tumors are important factors that are not categorical but are close to the circle in the circular plot above.

```
summary(res.mca)
```

Through the summary of this MCA model we concluded that through MCA we were able to reduce the dimensions by large number but since the data is quite big with each attribute having multiple factors we need to look into 133 dimensions for a 90% data variation. This is quite large and hence one need to look out for other techniques to analyze this data set.

The below plot between first two dimensions identify 10 most important attributes and categories that explan ~4% data variation (Dim1 and Dim2)

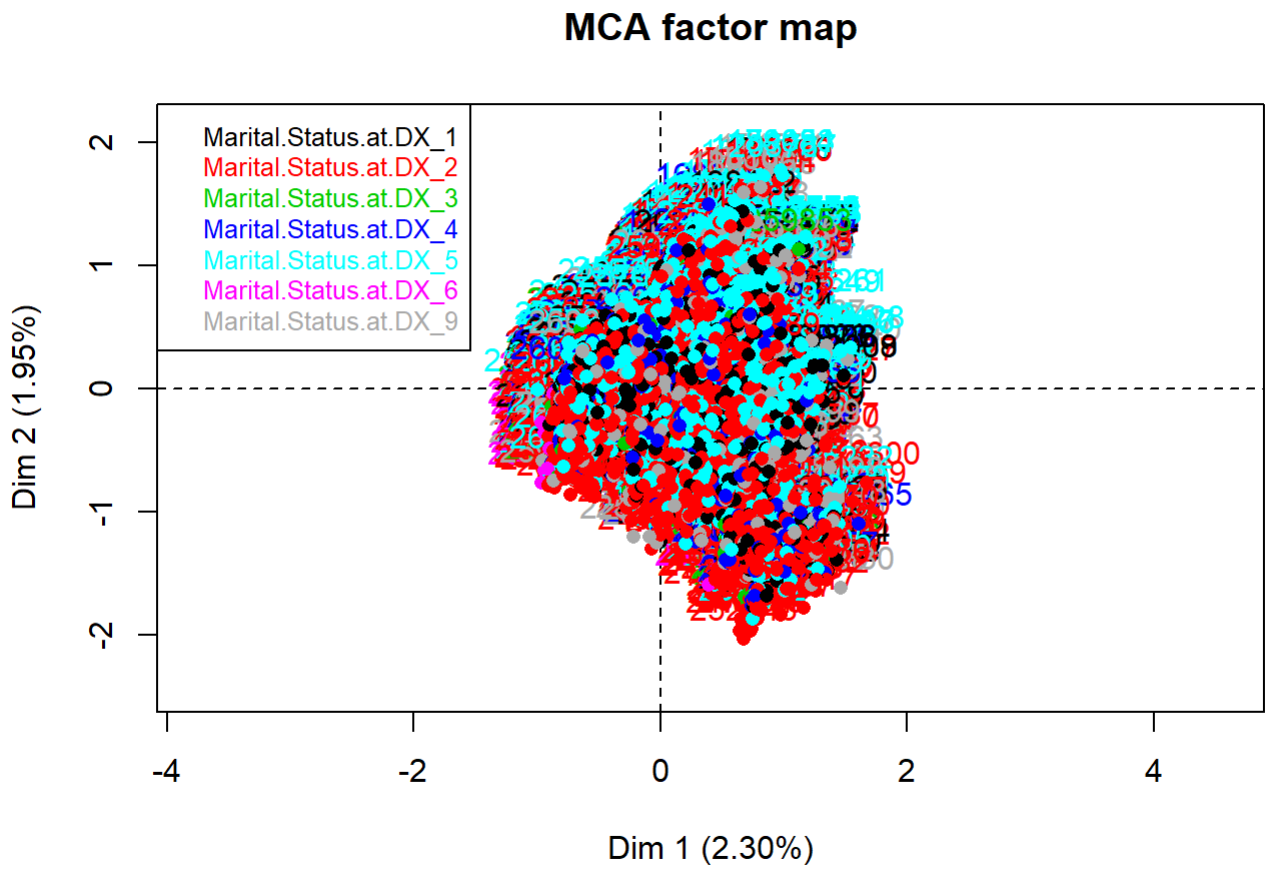


Form the above curve we identify the following as 10 most important attributes or categories in cancer analysis:

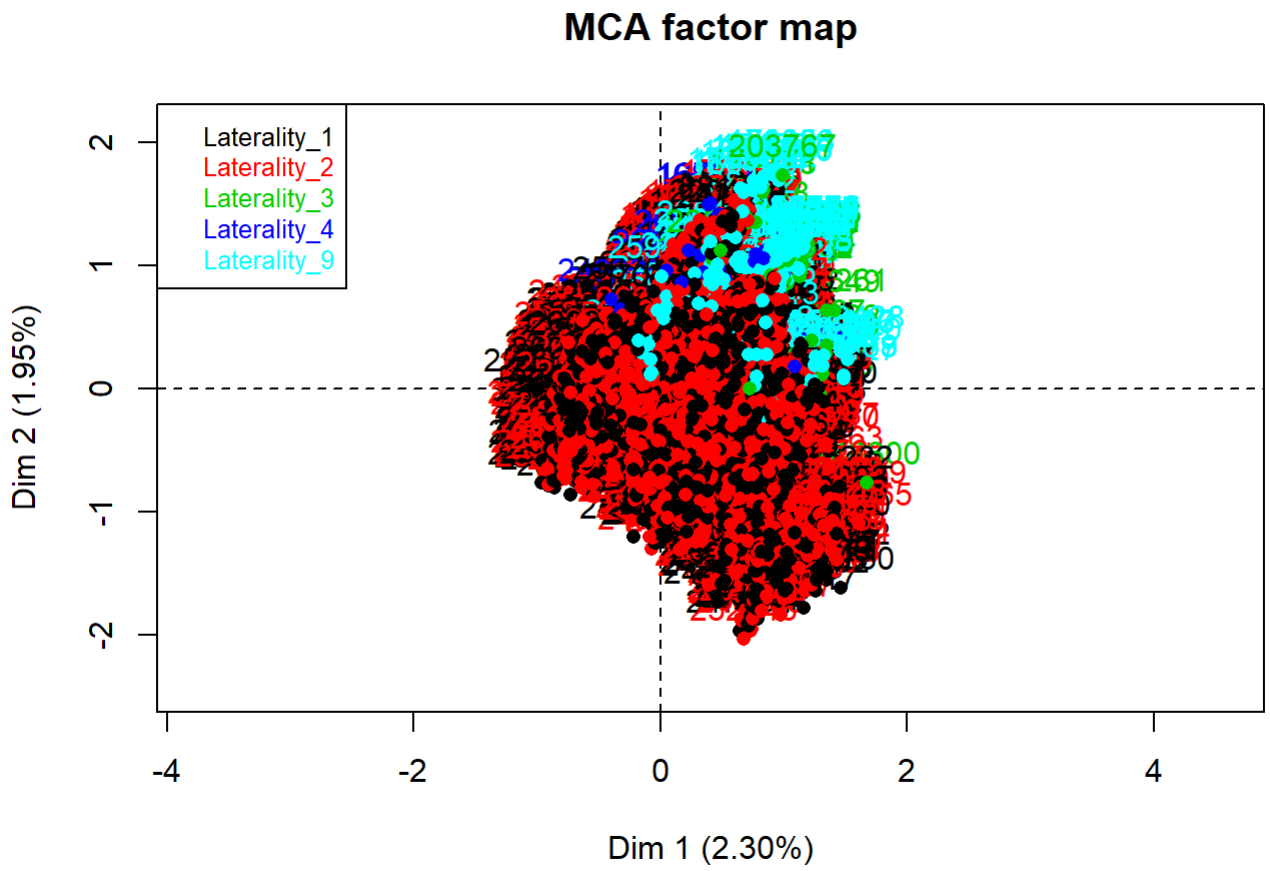
1. PR Status - 9,4
2. ER Status- 4
3. Grade -9
4. HER2 Status - 2
5. Breast Subtype - 3,5
6. Breast adjusted AJCC - 0
7. Behavior Code - 2

We also tried to plot the cancer cases by Marital status and laterality to identify any pattern

“MCA plot by Marital.Status”



"MCA plot by Laterality"

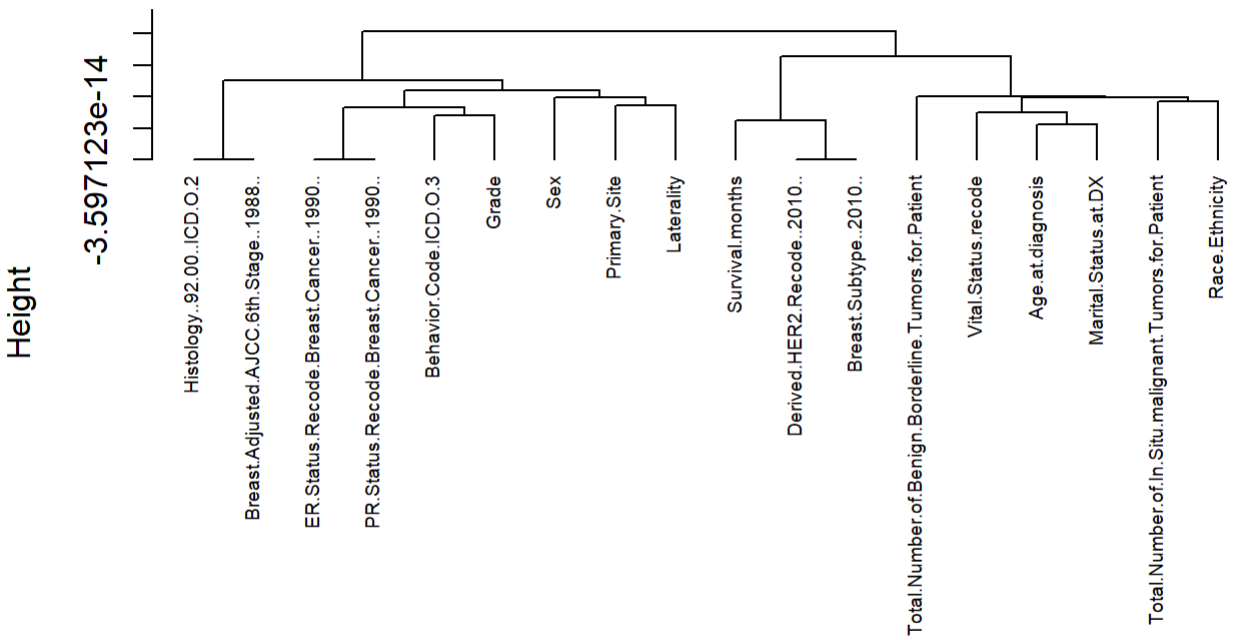


From the plot above the blue and the red group are two distinguish groups 2 - is for married and 4 - is for Divorced. Similarly, on the basis of laterality we have clear difference between bilateral and those with no laterality.

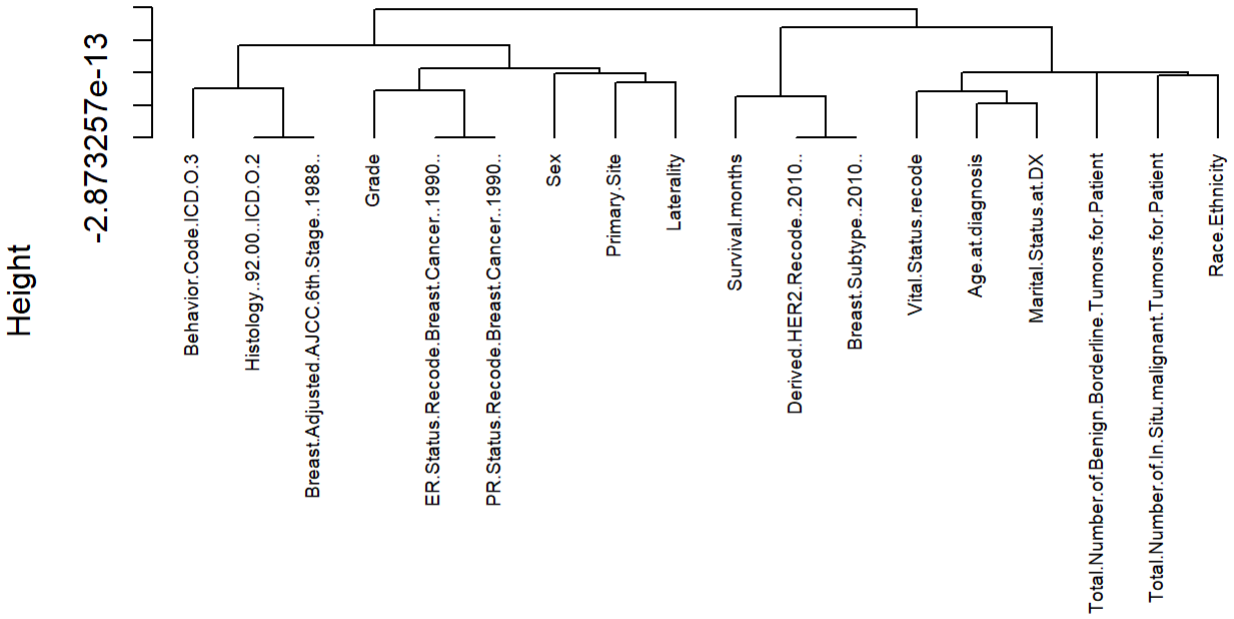
Apply clustering

To further analyze this dataset we applied the hierarchical clustering across different registry IDs to obtain dendrograms that cluster the attributes together. By this we will be able to compare the data distribution across different registries and find if the different areas have different cluster formations. This clustering can be implemented using a package called “ClustOfVar” in R. here are the various dendrograms obtained for 4 regions with most number of breast cancer cases.

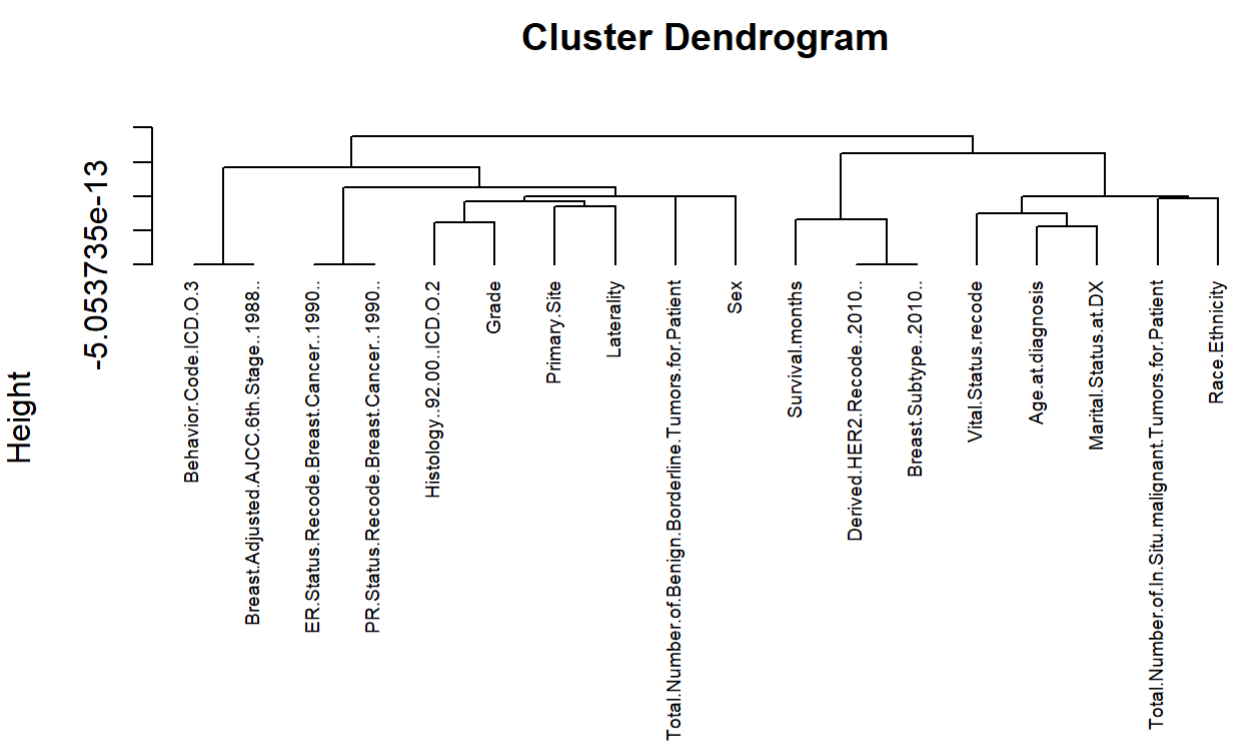
For 1501:


Cluster Dendrogram
Height
-3.597123e-14
Histology..92.00..ICD.O.2
Breast.Adjusted.AJCC.6th.Stage..1988..
ER.Status.Recode.Breast.Cancer..1990..
PR.Status.Recode.Breast.Cancer..1990..
Behavior.Code.ICD.O.3
Grade
Sex
Primary.Site
Laterality
Survival.months
Derived.HER2.Recode..2010..
Breast.Subtype..2010..
Total.Number.of.Benign.Borderline.Tumors.for.Patient
Vital.Status.recode
Age.at.diagnosis
Marital.Status.at.DX
Total.Number.of.In.Situ.malignant.Tumors.for.Patient
Race.Ethnicity

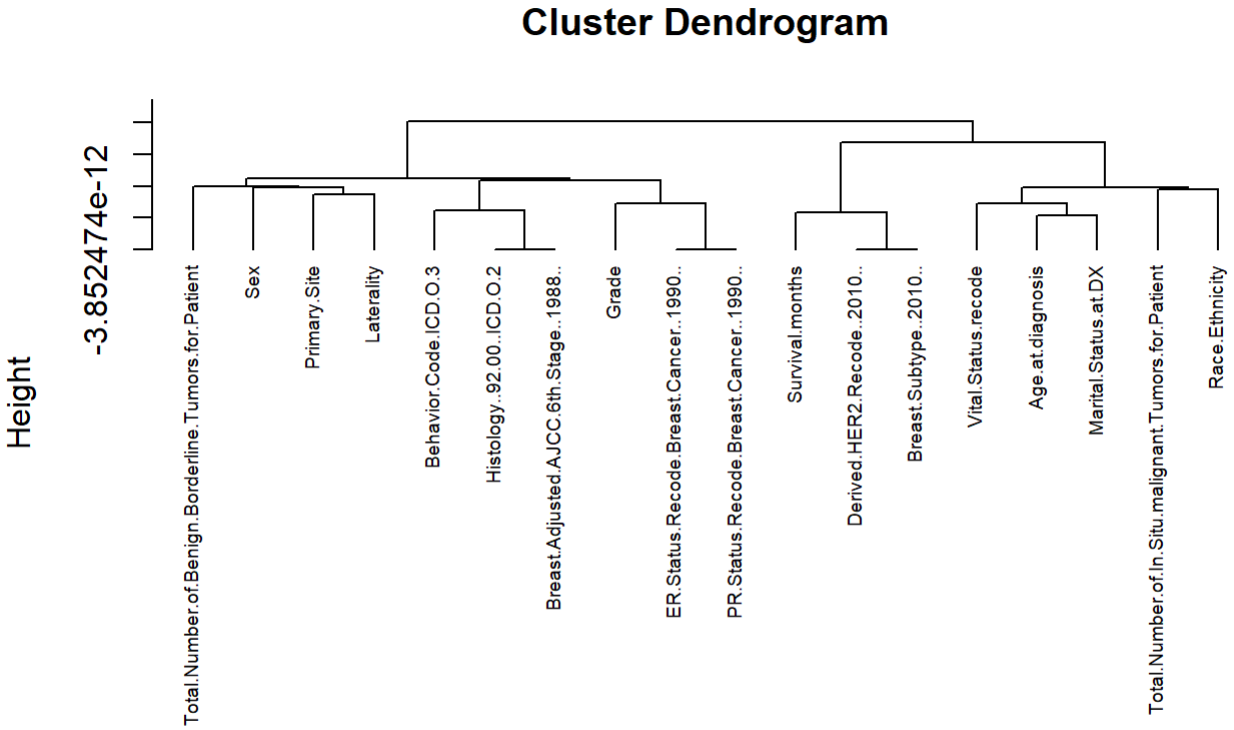
For 1502


Cluster Dendrogram
Height
-2.873257e-13
Behavior.Code.ICD.O.3
Histology..92.00..ICD.O.2
Breast.Adjusted.AJCC.6th.Stage..1988..
Grade
ER.Status.Recode.Breast.Cancer..1990..
PR.Status.Recode.Breast.Cancer..1990..
Sex
Primary.Site
Laterality
Survival.months
Derived.HER2.Recode..2010..
Breast.Subtype..2010..
Vital.Status.recode
Age.at.diagnosis
Marital.Status.at.DX
Total.Number.of.Benign.Borderline.Tumors.for.Patient
Total.Number.of.In.Situ.malignant.Tumors.for.Patient
Race.Ethnicity

For 1520


Cluster Dendrogram
Height
-5.053735e-13
Behavior.Code.ICD.O.3
Breast.Adjusted.AJCC.6th.Stage..1988..
ER.Status.Recode.Breast.Cancer..1990..
PR.Status.Recode.Breast.Cancer..1990..
Histology..92.00..ICD.O.2
Grade
Primary.Site
Laterality
Total.Number.of.Benign.Borderline.Tumors.for.Patient
Sex
Survival.months
Derived.HER2.Recode..2010..
Breast.Subtype..2010..
Vital.Status.recode
Age.at.diagnosis
Marital.Status.at.DX
Total.Number.of.In.Situ.malignant.Tumors.for.Patient
Race.Ethnicity

For 1525



The various dendrograms above can be used to cluster different attributes together. This help us in understanding how the different attributes are correlated to each other and if we can use it in dimension reduction or two study one attribute with respect to other. We then used stability curves and found that 6 is the optimum numbee of clusters for these dendrograms as above and hence following 6 Clusters were obtained:

Cluster-1: Race and Ethnicity, Number of malignant Tumors

Cluster-2: Marital Status, Age, Survival Status

Cluster-3: Survival months, breast subtype and HER2 status

Cluster-4: Laterality, primary site and sex

Cluster-5: Grade, behavior and PR status

Cluster-6: Histology, ER status and adjusted AJCC

Cox Proportional Hazard method

The clustering and the MCA analysis provide us with important attributes and their co-relations with each other, However to carry out the survival analysis and predict the probability of survival we need to used cox proportional hazard method and Kaplan-Meier estimator method. This method uses two output variables: 1. The survival flag and 2. Survival period after disease detection

All the other variables are used as input and the outcome is regression equation. We combine survival months and survival status as input to the cox model for regression. This is implemented using Surv() function from the “Survival Package” of R.

```
coxdata$SurvObj <- with(coxdata, Surv(coxdata$Survival.months, status == 1))
```

The cox proportional hazard regression method to identify the important parameters and hazards for the breast cancer patients is implemented using coxph() function as illustrated below:

```
## Fit Cox regression: age, sex, Karnofsky performance score, wt loss
res.cox1 <- coxph(SurvObj ~ Sex+
                  Marital.Status.at.DX+
                  Race.Ethnicity+
                  Laterality+
                  Histology..92.00..ICD.O.2+
                  Behavior.Code.ICD.O.3+
                  Grade+
                  ER.Status.Recode.Breast.Cancer..1990..+
                  PR.Status.Recode.Breast.Cancer..1990..+
                  Breast.Adjusted.AJCC.6th.Stage..1988..+
                  Breast.Subtype..2010..+
                  Age.at.diagnosis+
                  Total.Number.of.Benign.Borderline.Tumors.for.Patient+
                  Total.Number.of.In.Situ.malignant.Tumors.for.Patient, data =  coxdata)
```

Using the above code with survival object as output and other attributes as input we obtained following results and conclusions:

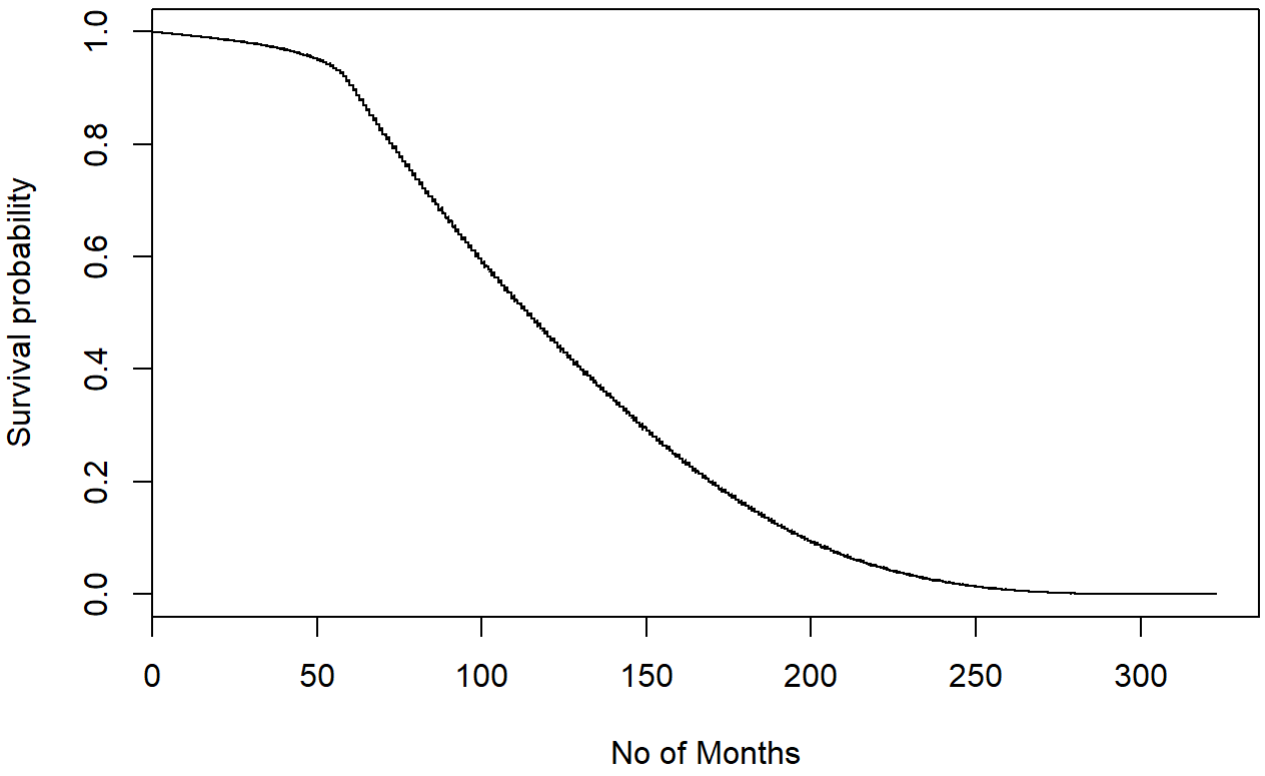
```
summary(res.cox1)
```

```
## Call:
## coxph(formula = SurvObj ~ Sex + Marital.Status.at.DX + Race.Ethnicity +
##     Laterality + Histology..92.00..ICD.O.2 + Behavior.Code.ICD.O.3 +
##     Grade + ER.Status.Recode.Breast.Cancer..1990.. + PR.Status.Recode.Breast.Cancer..1990.. +

##     Breast.Adjusted.AJCC.6th.Stage..1988.. + Breast.Subtype..2010.. +
##     Age.at.diagnosis + Total.Number.of.Benign.Borderline.Tumors.for.Patient +
##     Total.Number.of.In.Situ.malignant.Tumors.for.Patient, data = coxdata)
##
##   n= 619610, number of events= 397444
##
##                                                       coef  exp(coef)
## Sex                                                -9.053e-03  9.910e-01
## Marital.Status.at.DX                                9.745e-03  1.010e+00
## Race.Ethnicity                                      4.631e-03  1.005e+00
## Laterality                                          2.075e-02  1.021e+00
## Histology..92.00..ICD.O.2                          -1.458e-04  9.999e-01
## Behavior.Code.ICD.O.3                              -9.605e-01  3.827e-01
## Grade                                              -1.020e-01  9.030e-01
## ER.Status.Recode.Breast.Cancer..1990..             -2.111e-01  8.097e-01
## PR.Status.Recode.Breast.Cancer..1990..             -1.081e-01  8.975e-01
## Breast.Adjusted.AJCC.6th.Stage..1988..             -9.251e-04  9.991e-01
## Breast.Subtype..2010..                             -6.665e-01  5.135e-01
## Age.at.diagnosis                                   -3.752e-04  9.996e-01
## Total.Number.of.Benign.Borderline.Tumors.for.Patient -4.157e-02  9.593e-01
## Total.Number.of.In.Situ.malignant.Tumors.for.Patient -1.189e-01  8.879e-01
##                                                     se(coef)        z
## Sex                                                 2.300e-02   -0.394
## Marital.Status.at.DX                                9.606e-04   10.145
## Race.Ethnicity                                      1.514e-04   30.585
## Laterality                                          2.999e-03    6.918
## Histology..92.00..ICD.O.2                           1.947e-05   -7.490
## Behavior.Code.ICD.O.3                               4.826e-03 -199.010
## Grade                                               7.077e-04 -144.199
## ER.Status.Recode.Breast.Cancer..1990..              3.410e-03  -61.895
## PR.Status.Recode.Breast.Cancer..1990..              3.353e-03  -32.240
## Breast.Adjusted.AJCC.6th.Stage..1988..              8.061e-05  -11.476
## Breast.Subtype..2010..                              1.207e-03 -552.329
## Age.at.diagnosis                                    1.333e-04   -2.814
## Total.Number.of.Benign.Borderline.Tumors.for.Patient  2.159e-02   -1.926
## Total.Number.of.In.Situ.malignant.Tumors.for.Patient  2.618e-03  -45.403
##                                                    Pr(>|z|)
## Sex                                                 0.69387
## Marital.Status.at.DX                                < 2e-16 ***
## Race.Ethnicity                                      < 2e-16 ***
## Laterality                                         4.58e-12 ***
## Histology..92.00..ICD.O.2                          6.86e-14 ***
## Behavior.Code.ICD.O.3                               < 2e-16 ***
## Grade                                               < 2e-16 ***
## ER.Status.Recode.Breast.Cancer..1990..              < 2e-16 ***
## PR.Status.Recode.Breast.Cancer..1990..              < 2e-16 ***
## Breast.Adjusted.AJCC.6th.Stage..1988..              < 2e-16 ***
## Breast.Subtype..2010..                              < 2e-16 ***
```

```
## Age.at.diagnosis                                      0.00489 **
## Total.Number.of.Benign.Borderline.Tumors.for.Patient  0.05414 .
## Total.Number.of.In.Situ.malignant.Tumors.for.Patient  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                                      exp(coef) exp(-coef)
## Sex                                                     0.9910     1.0091
## Marital.Status.at.DX                                    1.0098     0.9903
## Race.Ethnicity                                          1.0046     0.9954
## Laterality                                              1.0210     0.9795
## Histology..92.00..ICD.O.2                               0.9999     1.0001
## Behavior.Code.ICD.O.3                                   0.3827     2.6129
## Grade                                                   0.9030     1.1074
## ER.Status.Recode.Breast.Cancer..1990..                  0.8097     1.2350
## PR.Status.Recode.Breast.Cancer..1990..                  0.8975     1.1142
## Breast.Adjusted.AJCC.6th.Stage..1988..                  0.9991     1.0009
## Breast.Subtype..2010..                                  0.5135     1.9474
## Age.at.diagnosis                                        0.9996     1.0004
## Total.Number.of.Benign.Borderline.Tumors.for.Patient    0.9593     1.0425
## Total.Number.of.In.Situ.malignant.Tumors.for.Patient    0.8879     1.1262
##                                                      lower .95 upper .95
## Sex                                                     0.9473    1.0367
## Marital.Status.at.DX                                    1.0079    1.0117
## Race.Ethnicity                                          1.0043    1.0049
## Laterality                                              1.0150    1.0270
## Histology..92.00..ICD.O.2                               0.9998    0.9999
## Behavior.Code.ICD.O.3                                   0.3791    0.3864
## Grade                                                   0.9017    0.9042
## ER.Status.Recode.Breast.Cancer..1990..                  0.8043    0.8151
## PR.Status.Recode.Breast.Cancer..1990..                  0.8917    0.9035
## Breast.Adjusted.AJCC.6th.Stage..1988..                  0.9989    0.9992
## Breast.Subtype..2010..                                  0.5123    0.5147
## Age.at.diagnosis                                        0.9994    0.9999
## Total.Number.of.Benign.Borderline.Tumors.for.Patient    0.9195    1.0007
## Total.Number.of.In.Situ.malignant.Tumors.for.Patient    0.8834    0.8925
##
## Concordance= 0.816  (se = 0.001 )
## Rsquare= 0.613   (max possible= 1 )
## Likelihood ratio test= 588711  on 14 df,   p=0
## Wald test            = 406410  on 14 df,   p=0
## Score (logrank) test = 918915  on 14 df,   p=0
```

We obtain a R2 value of 61% and concordance of 81%. The attributes with lower p values and positive coefficients are the most hazard causing elements that affect the survival negatively. From the above Cox proportional regression method we can say that race, histology, laterality and marital status are hazards to the survival as they have positive coefficients compared to others. We obtain an equation which consider all the other attributes as inputs. This cox proportional hazard equation can be used to predict the probability of survival for a given cancer case. Let us plot a survival fit curve that gives a plot between probability of survival and the number of months:



The plot gives a picture with low probability of a cancer patient surviving more than 100 months. This indicates that breast cancer is big hazard and pose a high probability of death.

Model Evaluation

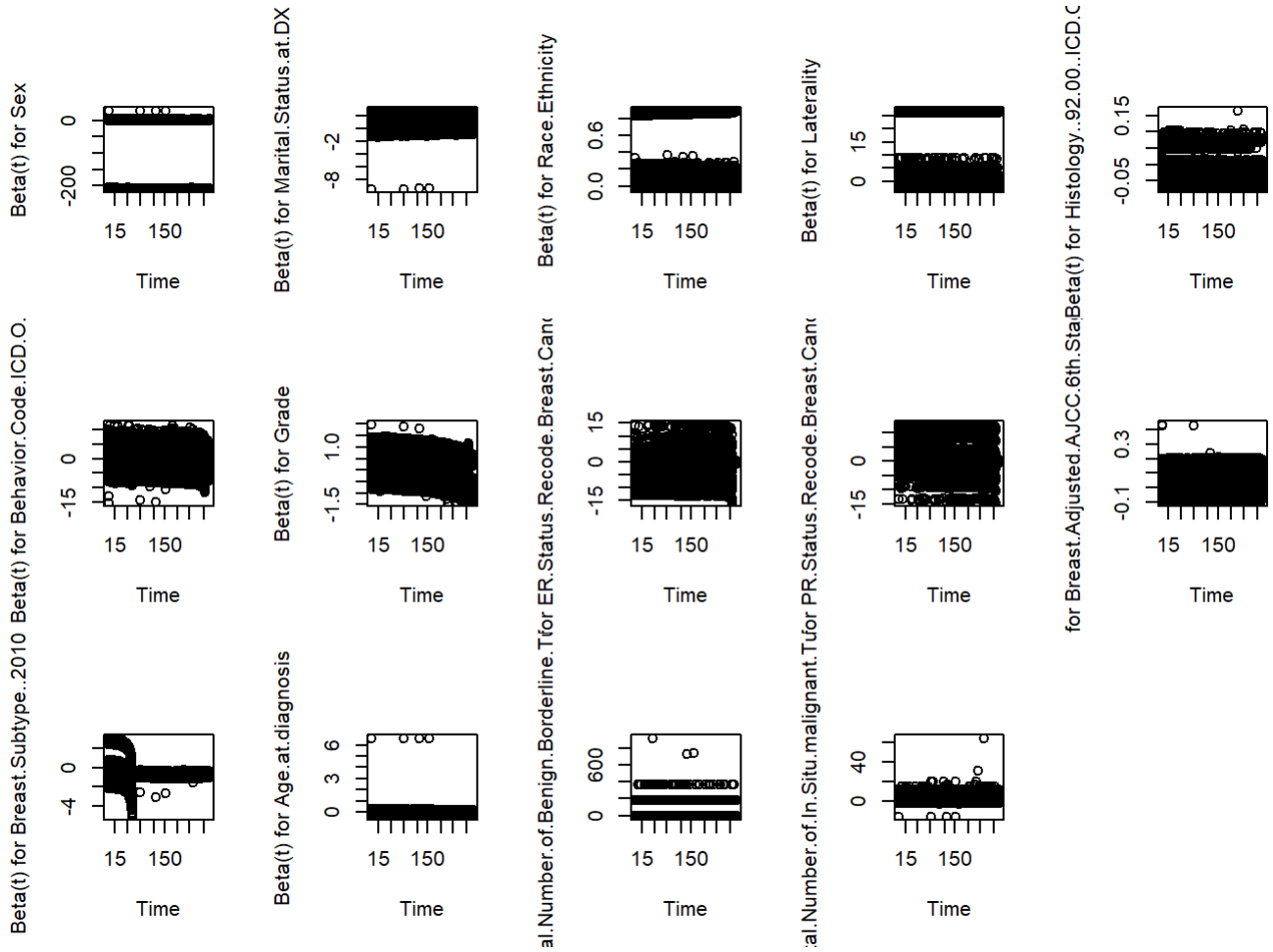
To further evaluate the performance of this cox proportional model we can test the proportional hazard assumption using the coz.zph() function. The proportional hazard assumption is supported by a non-significant relationship between residuals and time, and refuted by a significant relationship. The graphs below between the various time and beta value for various variables show that there is no relationship between time and residuals and hence the model is a good fit for the given dataset.

Note that the P values are low but since we have large dataset the P-values are low which indicate that cox regression may not be a better fit but in a large dataset p-values are often not the correct guide of model performance.

```
## Check for violation of proportional hazard (constant HR over time)
(res.zph1 <- cox.zph(res.cox1))
```

```
##                                                        rho    chisq
## Sex                                                -0.000757    0.228
## Marital.Status.at.DX                                0.001592    1.112
## Race.Ethnicity                                      0.009308   33.310
## Laterality                                         -0.000909    0.370
## Histology..92.00..ICD.O.2                           0.012485   70.627
## Behavior.Code.ICD.O.3                              -0.003512    4.748
## Grade                                              -0.017891  123.650
## ER.Status.Recode.Breast.Cancer..1990..             -0.006703   16.859
## PR.Status.Recode.Breast.Cancer..1990..             -0.007662   21.949
## Breast.Adjusted.AJCC.6th.Stage..1988..              0.021021  169.485
## Breast.Subtype..2010..                             -0.047862  520.776
## Age.at.diagnosis                                   -0.011421   54.139
## Total.Number.of.Benign.Borderline.Tumors.for.Patient  0.003493    4.834
## Total.Number.of.In.Situ.malignant.Tumors.for.Patient  0.000725    0.213
## GLOBAL                                                    NA 1962.728
##                                                           p
## Sex                                                8.33e-01
## Marital.Status.at.DX                               2.92e-01
## Race.Ethnicity                                     7.86e-09
## Laterality                                         5.43e-01
## Histology..92.00..ICD.O.2                          0.00e+00
## Behavior.Code.ICD.O.3                              2.93e-02
## Grade                                              0.00e+00
## ER.Status.Recode.Breast.Cancer..1990..             4.03e-05
## PR.Status.Recode.Breast.Cancer..1990..             2.80e-06
## Breast.Adjusted.AJCC.6th.Stage..1988..             0.00e+00
## Breast.Subtype..2010..                             0.00e+00
## Age.at.diagnosis                                   1.87e-13
## Total.Number.of.Benign.Borderline.Tumors.for.Patient 2.79e-02
## Total.Number.of.In.Situ.malignant.Tumors.for.Patient 6.44e-01
## GLOBAL                                             0.00e+00
```

```
par(mfrow=c(3,5))
plot(res.zph1)
```



Results

From the above analysis, we obtained following results:

1. Though the MCA we obtained most 10 most important attribute categories that affect the cancer patient's survival. These were the status of HER2, PR and ER cells that affect the health of the patients largely. Further the breast subtype, grade and behavior code were also identified as important attributes. These are attributes that define the cancer history, type of cancer cell and stage of cancer that affect the survival directly.
2. With further clustering following attribute clusters were identified:

Cluster-1: Race and Ethnicity, Number of malignant Tumors

Cluster-2: Marital Status, Age, Survival Status

Cluster-3: Survival months, breast subtype and HER2 status

Cluster-4: Laterality, primary site and sex

Cluster-5: Grade, behavior and PR status

Cluster-6: Histology, ER status and adjusted AJCC

6 clusters were identified using the Clustering in R and one use the Cox method to identify which cluster represent the group of attributes that affect the survival of cancer patient.

3. The Cox proportional hazard method further gave us the equation to predict the survival period of the cancer patients and also indicated that the attributes like the race, histology, laterality and marital status have more influence on the survival of a cancer patient as they have positive coefficients compared to others. The performance curves further strengthen this conclusions as there is no relationship found between the residuals and the time.

4. Thus with the three analyses that we conducted above we can say the following:

i. Breast cancer is a very common disease in women of all age groups and results in death for majority of the cases.

ii. It affects the married women more than the unmarried ones.

iii. The status of the cancer cells HER2, PR and ER affect the survival. The positive the status results in higher hazard.

iv. The survival probability decreases with increase in survival months indicating lower survival time period of cancer patients.

v. Race, histology or the patient history , laterality and marital status were the major hazards identified by cox regression method.

Discussion

The above analysis have helped us in analyzing the survival of breast cancer patients and identifying the various attributes that affect the survival at large. The cox proportional methods has given a fit curve and the MCA along with clustering has provided us with important attribute identification. However one can further identify the limitations of these methods and work out the better prediction algorithms.

The P-values in the test of violation of proportional hazard show that the cox regression may not be the best fit and hence one need to look out for other survival algorithms.

The MCA has alo proved to be not a very effective method for this large dataset and hence as a future scope to this project one should look out for other dimension reduction techniques which can be implemented on categorical data sets.

References

1. Dursun Delen, Analysis of cancer data: a data mining approach, Article in Expert Systems · February 2009

2. Dirk F. Moore, Applied Survival Analysis Using R, ISBN 978-3-319-31245-3

3. David Newitt, Nola Hylton, on behalf of the I-SPY 1 Network and ACRIN 6657 Trial Team. (2016). Multi-center breast DCE-MRI data and segmentations from patients in the I-SPY 1/ACRIN 6657 trials. The Cancer Imaging Archive. http://doi.org/10.7937/K9/TCIA.2016.HdHpgJLK (http://doi.org/10.7937/K9/TCIA.2016.HdHpgJLK)

4. Hylton NM, Gatsonis CA, Rosen MA, et al: Neoadjuvant Chemotherapy for Breast Cancer: Functional Tumor Volume by MR Imaging Predicts Recurrence-free Survival-Results from the ACRIN 6657/CALGB 150007 I-SPY 1 TRIAL. Radiology 279:44-55, 2016

5. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Phillips S, Maffitt D, Pringle M, Tarbox L, Prior F. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository, Journal of Digital Imaging, Volume 26, Number 6, December, 2013, pp 1045-1057.

6. Data clustering: algorithms and applications edited by charu c. aggarwal, chandan k. reddy

7. SEER (SURVEILLANCE, EPIDEMIOLOGY, AND END RESULTS) PROGRAM (2004) Public-Use Data (1973- 2001), National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, April; based on the November 2003 submission; www.seer.cancer.gov.

8. Data Clustering: Algorithms and Applications (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series) byCharu C. Aggarwal,Chandan K. Reddy

9.These websites provide a lot of codes and details for the cox regression, MCA and Clustering:

http://rstudio-pubs-static.s3.amazonaws.com/5896_8f0fed2ccbbd42489276e554a05af87e.html (http://rstudio-pubs-static.s3.amazonaws.com/5896_8f0fed2ccbbd42489276e554a05af87e.html)

http://dni-institute.in/blogs/cox-regression-interpret-result-and-predict/ (http://dni-institute.in/blogs/cox-regression-interpret-result-and-predict/)

https://www.r-bloggers.com/cox-model-assumptions/ (https://www.r-bloggers.com/cox-model-assumptions/)

https://seer.cancer.gov/data/ (https://seer.cancer.gov/data/)