"A statistical analysis of detecting different subtypes of breast Cancer "

Scientific Background:

The purpose of the study was to classify the subtypes of breast-cancer genes, whether two individuals with different breast cancer subtypes but the same age at diagnosis, tumour size and lymph node metastasis status will have different prognosis for survival.

Breast Cancer:

Breast cancer is very prevalent in woman globally, and the second most common death among women, after lung cancer. Breast Cancer remain one of the most common cancer in women worldwide and still prove difficult to treat effectively. There is evidence that different sub-types of breast cancer exist leading to the use of multi-modal therapy when the sub-types are unknown. Establishing breast cancer sub-types early would reduce treatment cost and also potentially increase treatment effectiveness.

Breast cancer is a type of cancer where the cells begin to grow out of control, most of the breast cancer are benign and not malignant. There are some types of benign breast cancer that can increase a women's risk of malignant with in a future event.

BRCA1(BReast CAncer gene) and BRCA2 are two different genes that have been found to impact a person's chance of developing breast cancer. There are five main intrinsic or molecular subtypes of breast cancer that are based on the genes a cancer expresses:

- Luminal A
- Luminal B
- Triple-negative/basal-like
- HER2-enriched
- Normal-Like

To begin our analysis, we have data collected of single-channel microarray gene expression data from 251 patients presenting with breast cancer. The gene expression data are (unlogged) intensity values from single-channel microarray technology. We conduct several statistical tests to determine the subtypes and check our analysis with our areas of interest.

Methods:

The statistical methods that I had conducted during our analysis were

1. Cluster Analyses
2. Silhouette method for finding the optimal number of clusters
3. Principal component analysis
4. Differential Gene Analysis
5. Survival Analysis

From the above-mentioned statistical procedures, we conduct our analysis to find the subtypes of the breast cancer genes.

Cluster Analysis:

Cluster analysis divides data into groups that are both meaningful and useful. In our current analysis, we use hierarchical clustering technique.

Hierarchical Clustering: A set of nested clusters organised as a hierarchical tree. They can be represented using traditional clusters or dendrograms.

Two main type of hierarchical clustering are:

- Agglomerative
- Divisive

The distance between the clusters is measured using Euclidean distance:

d (A, B) = $\sqrt{\sum_{i=1}^{k}(A_i - B_i)^2}$

Between, two points A and B with k dimensions.

We tend to analyse our micro-array data using the clustering technique and cut the dendrogram at a limit. However, we would like to know the optimal number for clusters for our sample set. For determining the optimal number of clusters, we use the silhouette method.

Silhouette method:

$s_i = \frac{b_i - a_i}{max(a_i, b_i)} \in [0,1]$, where a_i= $avg\ distanc$e of the sample 'i' to the other sample that belongs to the same cluster; b_i= $avg\ distanc$e of the sample 'i' and all sample that it doesn't belongs to.

Principal Component Analysis:

PCA is a very efficient technique used to de-noise and find patterns in large datasets and is generally used as a dimensionality reduction method which allows us to visualize high-dimensional data.

Let S = $\frac{1}{n-1}\sum_{i=1}(x_i - \bar{x})(x_i - \bar{x})^T$ be the sample variance-covariance matrix

$\bar{x}$ is the sample mean that contains the mean of each of the k feature of the vectors x_i. We can make the eigen-decomposition.

S = UDU^T, where U is n x k orthogonal matrix containing the (unit) eigen-vectors of S and D is a k x k diagonal matrix containing the eigenvalues of S.

Firstly, the eigenvectors of the variance-covariance matrix are calculated, the vectors determine the directions of

maximum variance which are known as principal components. The eigenvalues are then created to determine the value of the principal components (i.e.) the eigenvalues are the principal components.

A key point that has to be noted here is that eigenvectors with largest eigenvalues are the ones with highest dispersion from the mean and they are closest to the original data set. Also, larger the number of eigenvectors, the computation performance is slower. Thus, where our data comes from higher dimension, we perform PCA to preserve as much as distance between the samples as possible.

Differential Gene Analysis:

We perform DGE analysis to determine which genes are expressed at various levels between two experimental conditions., these genes can offer greater insights into the process affected by the condition(s) of interest and to understand their level of statistical significance.

Before conducting DGE analysis it is good to omit genes that have little or no chance of being detected as differentially expressed.

Now that we have our clusters determined we can get the subtypes of the data from the dendrograms, by comparing the clusters and obtaining a list of Differentially expressed genes with q-values < = 0.05.

We can conduct this by doing an empirical bayes to find the prior distribution and the respective qvalue. Another method to find the DE genes is by using a simple t-test and conduct hypothesis tests stating our null and alternative hypothesis.

Survival Analysis:

Survival analysis analyses the expected duration of time until a particular or multiple event occurs such as development of malignant tumour among cancer patients or probability of failure of a manufacturing equipment based on hours of operation.

F*(t)* = Pr(T $\leq t$) (F*(t)* probability of the event occurring). The survival function by,

S (t) = 1 − F (T)

= Pr (T > t),

Which is the probability that the event of interest may occur in the future. Next, let f (t) be the accompanying probability density function (pdf) that is:

f (t) = d F(t) / dt

is the derivative with respect to t, then we call

$$\lambda(t) = \frac{f\ (t)}{S\ (t)}$$

Where, $\lambda(t)$ = hazard function,

which is also known as instantaneous failure rate, it measures the prosperity of the event of interest occurring exactly at time t.

In proportional hazards model, $\lambda_1(t) = \theta\lambda_0(t)$, θ is hazard ratio parameter when comparing two hazard functions in a coherent study.

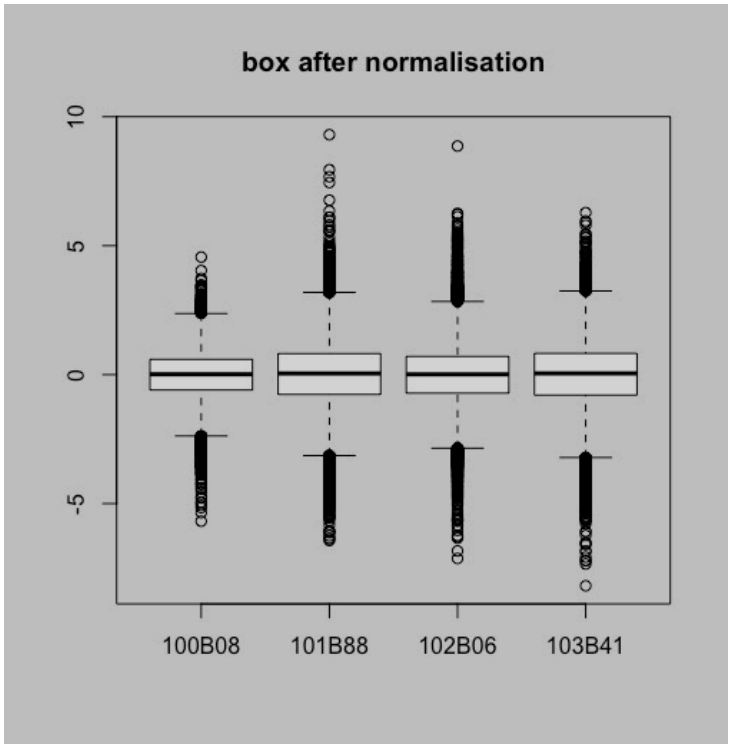
When θ>1, exposure increases your risk of event ⟹ decreasing survival rates

When θ<1, exposure decreases your risk of event ⟹ increasing survival rates

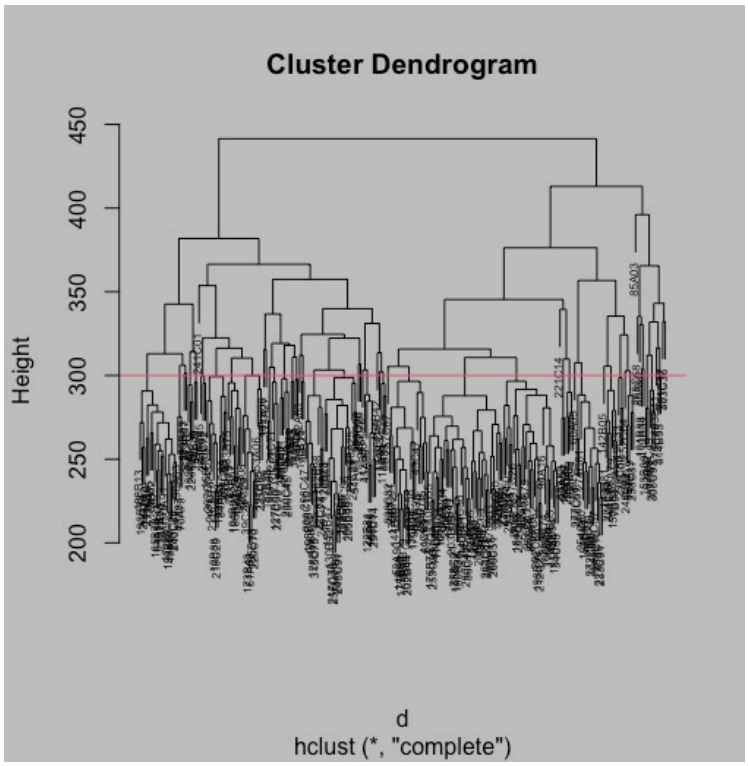
From these we can determine the event to occur in our study and take necessary precautions to treat the breast cancer.

<u>Results:</u>

Box plot after normalisation:



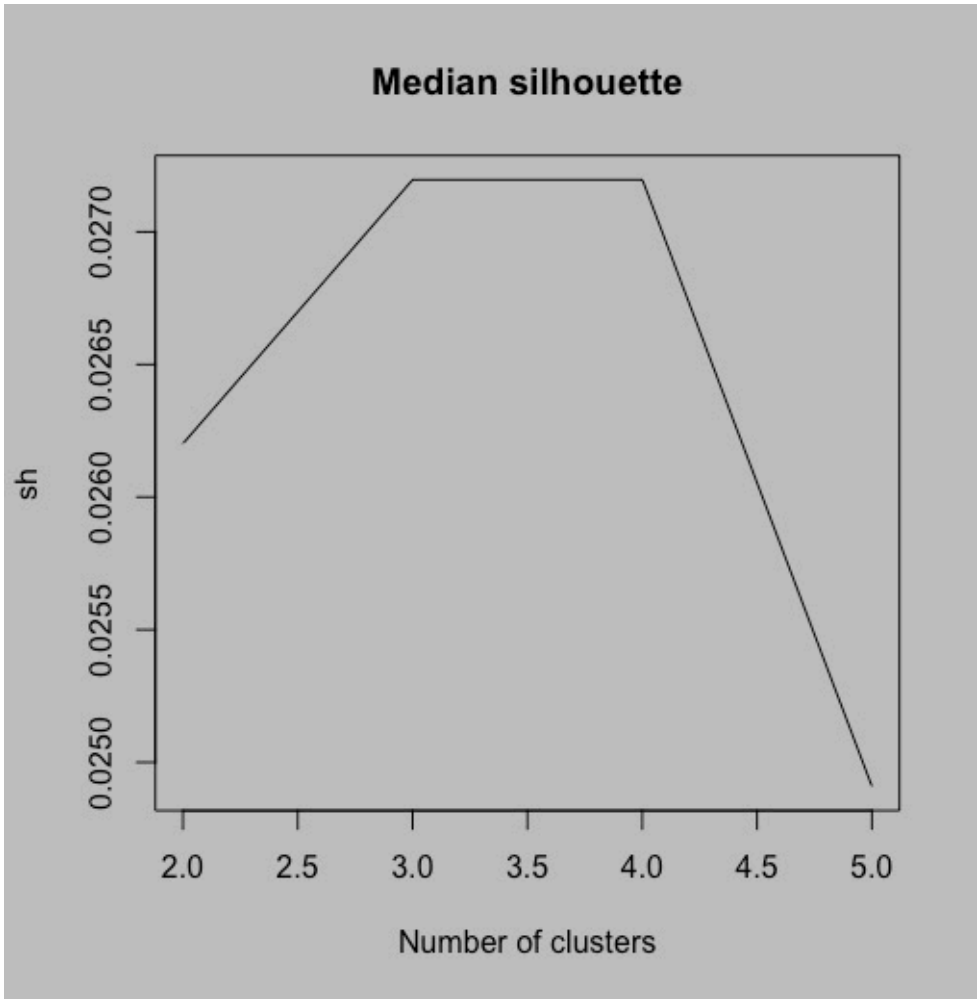
Clustering Result: In the clustering plot we add **abline**() which is used to add horizontal or regression lines to a graph which can be indicated by a red line.In the clustering we use the complete method to ensure that all datapoints are being covered by the clusters.



After obtaining our clusters we would be interested to know the ideal number of clusters for our datapoints. From the below figure we can deduce that if optimal number of clusters that cover all our datapoints are between 3 and 4.

Based on our optimal number of clusters we cut our dendrogram tree.

#Plot silhouette:

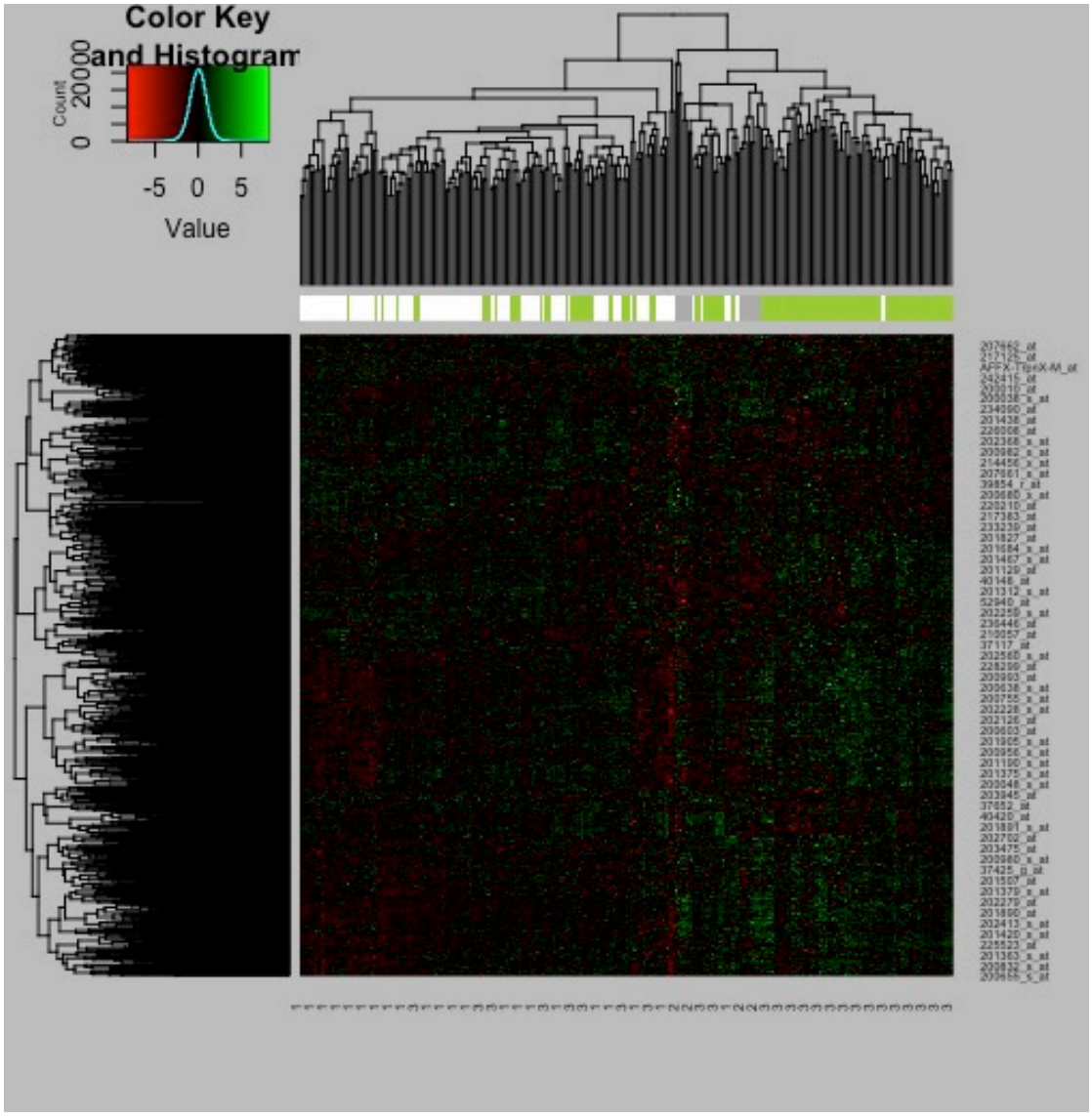



```
cl
  1   2   3
118  15 118
```

We have three cluster and the datapoints covered by them are given above.

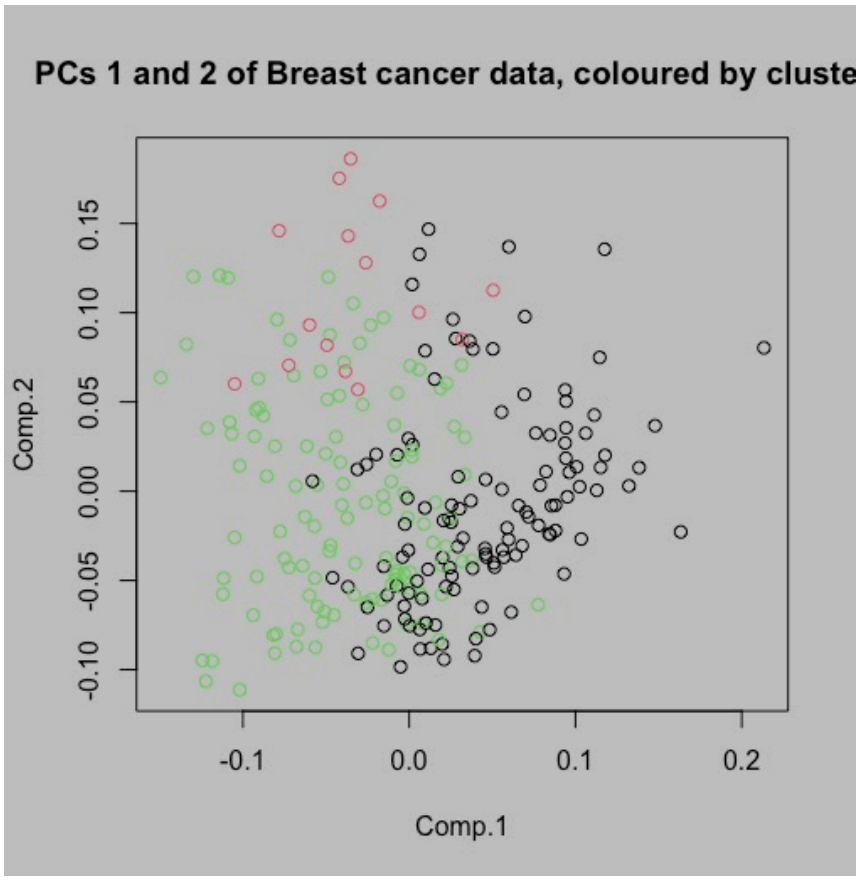
Heatmap:

A heat map is a data visualisation technique that represents the magnitude of a phenomenon as colour in two dimensions. The variation in colour may be by hue or intensity, giving obvious visual cues to the reader about how the phenomenon is clustered or varies over space.



PCA Analysis:

We further conduct PCA Analysis to determine the principal components and we perform PCA to preserve as much as distance between the samples as possible.



The three principal components are displayed by their respective colours which are green, red and black.

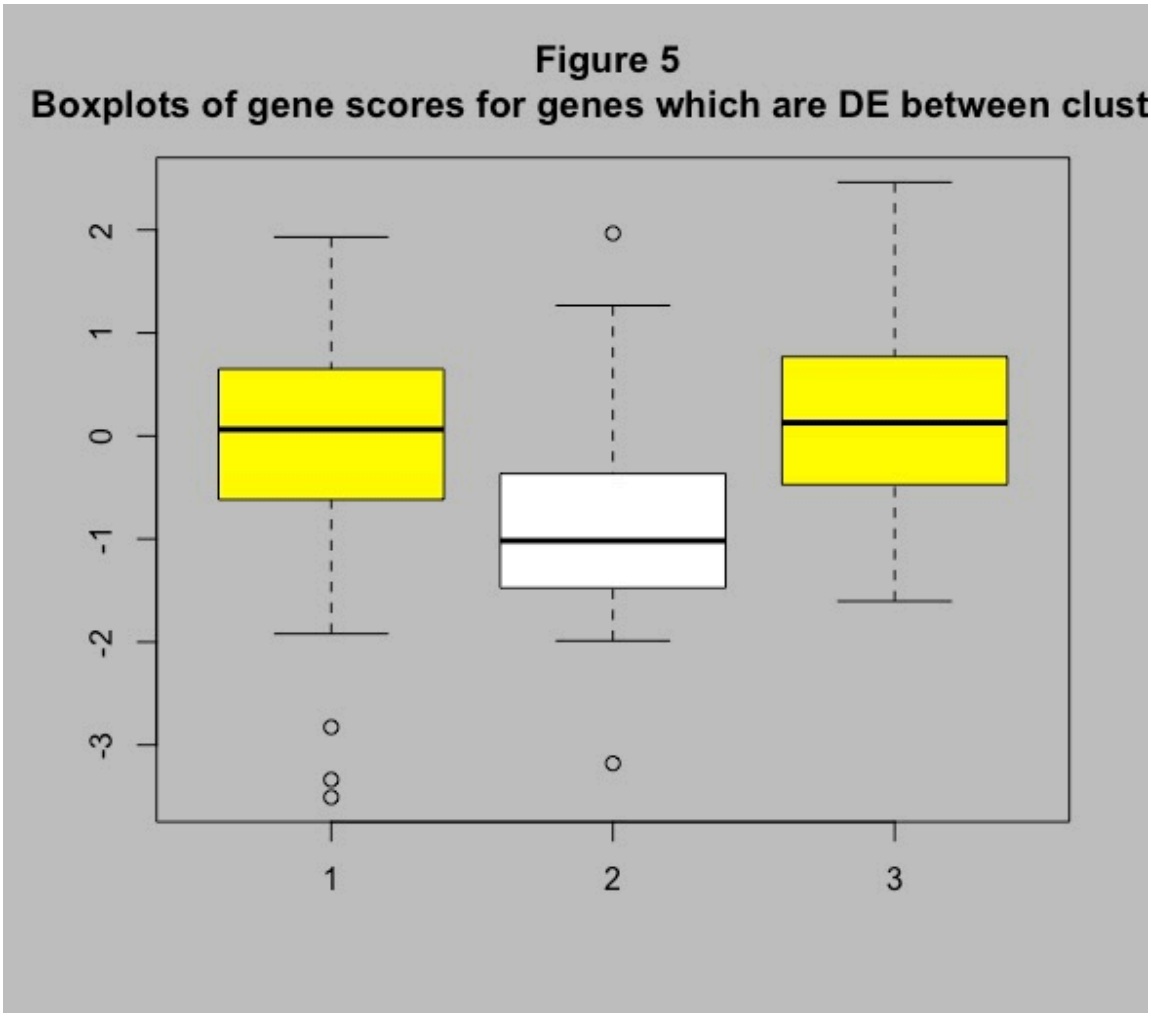
From the principal components we would be interested for further analysis where we can the DGE list from the components.

Survival Analysis:

We further do the DGE Analysis to determine the Obtain DE genes with qvalue<=0.05.

After generating the DGE genes with q-values less than 0.05, we form the gene expression based on the significant DE genes and standardise the scale with ($\mu = 0$ & $\sigma = 1$).

We further perform the Cox regression to estimate HR of gene.score, the score of each sample is the sum of the differentially expressed gene values for that sample.



If we can split the gene scores values by cluster number, we can visualise the gene scores for the DE genes between the three different clusters using a boxplot

A better summary can be viewed by the summary(cox.model) that will give the coefficient estimate is not statistically significantly different from 0 (for any $\alpha < 0.174$), so we cannot conclude that the hazard ratio is different from 1. This conclusion is supported by the 95% CI for the hazard ratio, which contains the value 1.

- `exp(-coef)` is therefore the (inverse) hazard ratio
- `coef` is this estimated coefficient ββ^ from the model
- `se(coef)` is the standard error
- `z` is the z-score
- `Pr(>|z|)` the propability estimated.
- `lower .95` and `upper .95` are the 95%-confidence interval for the estimated hazard ratio `exp(coef)`

```
> summary(cox.model)
Call:
coxph(formula = Surv(Surv_time, event) ~ histgrade + gene.score +
    ERstatus + PRstatus + tumor_size_mm + age + LNstatus, data = x)

  n= 236, number of events= 55
   (15 observations deleted due to missingness)

                    coef  exp(coef)   se(coef)      z Pr(>|z|)
histgradeG1   -1.772e+00  1.699e-01  1.086e+00 -1.632  0.10257
histgradeG2   -1.236e+00  2.905e-01  1.033e+00 -1.197  0.23133
histgradeG3   -1.042e+00  3.527e-01  1.085e+00 -0.961  0.33680
gene.score     1.910e-01  1.210e+00  1.364e-01  1.400  0.16147
ERstatusER?   -1.548e+01  1.891e-07  3.799e+03 -0.004  0.99675
ERstatusER+    6.438e-01  1.904e+00  5.510e-01  1.168  0.24270
PRstatusPgR+  -4.650e-01  6.282e-01  4.485e-01 -1.037  0.29991
tumor_size_mm  3.217e-02  1.033e+00  1.226e-02  2.624  0.00870 **
age            3.780e-03  1.004e+00  1.041e-02  0.363  0.71667
LNstatusLN?   -1.617e+01  9.495e-08  3.720e+03 -0.004  0.99653
LNstatusLN+    9.853e-01  2.679e+00  3.070e-01  3.209  0.00133 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

              exp(coef) exp(-coef) lower .95 upper .95
histgradeG1   1.699e-01  5.884e+00   0.02024     1.427
histgradeG2   2.905e-01  3.442e+00   0.03839     2.199
histgradeG3   3.527e-01  2.836e+00   0.04205     2.958
gene.score    1.210e+00  8.261e-01   0.92647     1.582
ERstatusER?   1.891e-07  5.288e+06   0.00000       Inf
ERstatusER+   1.904e+00  5.253e-01   0.64645     5.606
PRstatusPgR+  6.282e-01  1.592e+00   0.26079     1.513
tumor_size_mm 1.033e+00  9.683e-01   1.00817     1.058
age           1.004e+00  9.962e-01   0.98350     1.024
LNstatusLN?   9.495e-08  1.053e+07   0.00000       Inf
LNstatusLN+   2.679e+00  3.733e-01   1.46755     4.889

Concordance= 0.78  (se = 0.029 )
Likelihood ratio test= 46.22  on 11 df,   p=3e-06
Wald test            = 44.18  on 11 df,   p=7e-06
Score (logrank) test = 54.46  on 11 df,   p=1e-07
```

Conclusion:

From our analysis we have found few statically significant genes:

tumor_size_mm

coef:3.217e-02
exp(coef):1.033e+00
se(coef): 1.226e-02
z:2.624
Pr(>|z|):0.00870

A recent study [1] to determine the type of relationship between tumour size and mortality in early breast carcinoma. Females with severe tumour size had high mortality rate and the clinical importance of tumour size, and the functional form linking size to outcome.

LNstatusLN+

coef:9.853e-01
exp(coef):2.679e+00
se(coef): 3.070e-01
z: 3.209
Pr(>|z|):0.00133

The more lymph nodes that contain cancer cells, the more serious the cancer might be. This information, along with your cancer size and grade, are important clinical factors in deciding what treatment to pursue after surgery.

Out analysis has shown to be capable of understanding the risk of death due to breast cancer from data on the size of the tumour and the number of positive lymph nodes. In addition, this method was used to classify women into groups according to breast carcinoma severity. In contrast, classification of women according to lymph node positivity, tumour size, or disease stage created cluster groups with wide and mixed levels of lethality.

If such information can be determined, the required precautions and measures can be taken to ensure that the cancer does not become malignant in the further stages

References

1. Verschraegen C, Vinh-Hung V, Cserni G, et al. Modeling the effect of tumor size in early breast cancer. *Ann Surg*. 2005;241(2):309-318. doi:10.1097/01.sla.0000150245.45558.a9
2. Smith BL. Approaches to breast-cancer staging. *N Engl J Med*. 2000;342:580–581.
3. Cote RJ, Peterson HF, Chaiwun B, et al. Role of immunohis-tochemical detection of lymph-node metastases in manage- ment of breast cancer. International Breast Cancer Study Group. *Lancet*. 1999;354:896 –900.
4. https://www.mybreastcancertreatment.org/en-US/PersonalizeYourTreatment/Understanding-Node-Positive-Breast-Cancer