Exploring the Factors Influencing Breast Cancer Outcomes

By Shrinidhi Rajesh

03/16/2023

Abstract:

Breast cancer is a complex disease that presents significant challenges in understanding its progression and identifying factors that may influence outcomes. This study addresses these challenges by performing three different analysis. The main goal of this research is to perform a survival analysis on a breast cancer dataset to identify factors that influence the time until death, to investigate factors responsible for the likelihood of mortality, and to analyse the predictors that influence the increase in the number of lymph nodes The dataset comprised of 2982 primary breast cancer patients. The Cox proportional hazard model, logistic regression, and Zero Inflated Negative Binomial models were used for the respective analysis. The results showed that grade and recurrence were influencing factors for time until death, and age, grade, node, recurrence and hormonal treatment were responsible for the likelihood of mortality. The regression for count data analysis found that the Zero Inflated Negative Binomial model was the best fit for predicting the increase in the number of lymph nodes affected by tumor. Overall, this study provides valuable insights into breast cancer prognosis and identifying important predictors for better patient management.

Table of Contents:

Introduction:

Breast cancer is a complex and heterogeneous disease that poses significant challenges in understanding its prognosis. Prognosis refers to the likelihood of survival or disease recurrence after diagnosis and treatment. Factors that influence breast cancer prognosis include tumor characteristics such as size, grade, hormone receptor status, HER2 status, lymph node involvement, and molecular subtype, as well as patient characteristics such as age, comorbidities, and treatment history.

Several studies have investigated the factors that influence breast cancer prognosis. For instance, studies have shown that larger tumor size, higher tumor grade, lymph node involvement, and HER2-positive status are associated with a poorer prognosis. Moreover, younger age at diagnosis, higher BMI, and certain comorbidities such as diabetes and hypertension have been linked to a worse prognosis in breast cancer patients.

In addition to tumor and patient characteristics, treatment modalities also play a crucial role in breast cancer prognosis. Treatment options for breast cancer may include surgery, chemotherapy, radiation therapy, targeted therapy, and hormone therapy. The optimal treatment approach depends on several factors, including tumor characteristics, patient preferences, and the presence of comorbidities.

Overall, understanding the prognosis of breast cancer requires a comprehensive assessment of multiple factors that influence disease outcomes. Identifying prognostic factors can help guide treatment decisions and improve patient outcomes.

In this report, we analyze a dataset comprising of 2982 primary breast cancer patients to identify factors that influence time until death, investigate factors responsible for the likelihood of mortality, and analyze the predictors that influence the increase in the number of lymph nodes affected by tumor. We use three different statistical models, including the Cox proportional hazard model, logistic regression, and Zero Inflated Negative Binomial models, to perform our analyses.

The purpose of this report is to provide valuable insights into breast cancer prognosis and identify important predictors for better patient management. By understanding the factors that influence disease outcomes, healthcare professionals can guide treatment decisions and improve patient outcomes. The findings of this report may also contribute to the development of more effective and personalized treatment approaches for breast cancer.

Methods

The breast cancer dataset was obtained from Survival Package in R. The dataset named "Rotterdam" was written into csv from R and loaded into SAS. SAS software was used in this study, the robustness of this tool aided in performing statistical analyses and provided a high degree of flexibility and customization in data manipulation and reporting.
The Rotterdam data set includes 2982 primary breast cancers patients found in the Rotterdam tumour bank. The dataset includes patient data between the years 1978 and 1993**. The dataset comprised of** 15 variables.

pid	patient identifier
year	year of surgery
age	age at surgery
meno	menopausal status (0= premenopausal, 1= postmenopausal)
size	tumour size, a factor with levels <=20 20-50 >50
grade	differentiation grade fo Tumour
nodes	number of positive lymph nodes
pgr	progesterone receptors (fmol/l)
er	estrogen receptors (fmol/l)
hormon	hormonal treatment (0=no, 1=yes)
chemo	chemotherapy
rtime	days to relapse or last follow-up
recur	0= no relapse, 1= relapse
dtime	days to death or last follow-up
death	0= alive, 1= dead

In this dataset, the size variable consisted of values '<=20','20-50' and '>50', these three values were used to create a new variable called tumor_sze with dummy values, 0,1 and 2 respectively for better usage in the model. The analysis is done using SAS software throughout the study. The below is a detailed description of the column variables [Table 1].

Table 1: Variables and their measurements

Variable	Level of Measurement	Type of Data
Year	**Ratio**	**Continuous**
Age	**Ratio**	**Continuous**
Meno	**Nominal**	**Categorical**
Size	**Ordinal**	**Categorical**
Grade	**Ordinal**	**Categorical**
Nodes	**Ratio**	**Count**
Pgr	**Ratio**	**Continuous**
Er	**Ratio**	**Continuous**
hormon	**Nominal**	**Binary**
Chemo	**Nominal**	**Binary**
Rtime	**Ratio**	**Continuous**
Recur	**Nominal**	**Binary**
Dtime	**Ratio**	**Continuous**
Death	**Nominal**	**Binary**
Tumor_size	**Ordinal**	**Categorical**
Age_Catergory	**Ordinal**	**Categorical**

The loaded data was labelled and analysed for missing values [table 2].

Table 2: Checking for missing values

The MEANS Procedure

Variable	Label	N Miss
year	Year the patient underwent surgery	0
age	Age of Patient during surgery	0
meno	Menopausal Status (0=Premenopausal, 1=Postmenopausal)	0
tumor_size	Size of tumor, <=20 is 1, 20-50 is 2 and >50 is 3	0
grade	Grade of Tumor	0
nodes	Number of Positive Lymph Nodes	0
pgr	Progesterone Receptors(fmol/l)	0
er	Estrogen Receptors (fmol/l)	0
hormon	Hormonal Therapy (0=Not Received, 1=Received)	0
chemo	Chemotherapy received or not (0=Not Received, 1=Received)	0
rtime	days to relapse or last follow-up)	0
recur	Whether or not the patient experienced recurrence(0= no relapse, 1= relapse)	0
dtime	days to relapse or last follow-up	0
death	Whether the patient died during the follow-up period 0=Alive, 1=Death	0

With no missing values in the data, the descriptive statistics were performed on age [table 3]. The frequency distribution was calculated for nodes for better understanding of the number of affected lymph nodes and is found in [table 4]. The variables rtime and dtime were both on time to event thus their correlation was checked using correlation matrix [table 5]. With their correlation being high only the variable dtime was used in the forthcoming models. Frequency distributions were calculated for variable recur, chemo and hormon with death separately to understand if the patients with recurrence of breast cancer is facing higher death rate and similarly if chemo and hormon therapy the death [table 6].

The main goal of this study was to perform a survival analysis on the breast cancer dataset, with the aim of identifying factors that influence the time until mortality of the patient. For this Cox proportional hazard model was used using proc phreg procedure in SAS. The model used dtime as the response variable i.e the time till death or last followup and the variables age, meno, nodes, chemo, hormon, recur, grade as the predictors. The variables tumor_size was excluded as the column doesn't add any meaning to this analysis, rtime was not included as one amongst dtime and rtime variables would be enough. The variable death is excluded as it wont be adding any meaning to the analysis done here. The procedure phreg is used here. The Cox Proportional Hazard model was performed using stepwise selection method with the entry range as slentry=0.20 and final range as slstay=0.15 [table 7]. The result showed grade and recur to be an influencing factor which is utilized for the further representation, the Sruvival Curve was plotted used Lifetest procedure in SAS [table 8]

Secondly, I investigated the factors that are responsible for the likelihood of mortality. To perform this I used Logistic Regression as the response variable used in the prediction follows a binomial distribution, death=0,1. The predictors, age, meno, grade, node, recur, dtime, recur, hormon and chemo were used in the modelling. Tumor_grade and rtime were not considered here for the same reason as for Cox Proportional Model. Logistic Regression was performed using proc logistic and selection stepwise was used[table9]. To perform logistic regression, the assumption of independence should not be violated in the data. To check for the assumption of independence proc scatterplot was used with observation vs residual to perform the plot [table 10].

Lastly, I analysed which predictors acts crucial in influencing the increase in the number of lymph nodes affect by tumor cells causing metastasis. Nodes is a count variable, thus Regressions for count data is utilized for the prediction. Poisson, Negative Binomial, Zero Inflated Negative Binomial and Zero Inflated Poisson were performed for the comparisons to choose the best fit model. Firstly, the number of zeros in the response variable was calculated using proc sql [table 11], which showed there were 1436 zeros in the response variable which is more than 50% of the data and is really high, thus it is wise to go with Zero Inflated Negative Binomial or Zero Inflated Poisson. To reconfirm, Poisson

regression was run using Genmod procedure in sas with dist as Poisson and link function as log[table 12], the dispersion parameter value is 3.96 which clearly states overdispersion of the data, further calculating the mean and variation[table 13], the Variance value 19.22 was much larger than mean 2.71. Accounting the overdispersion, negative binomial was performed using Genmod, dist as NegBin and link as log [table 14]. Similarly Zero Inflated Negative Binomial [table 15] was performed with Genmod Procedure dist as ZINB, log as link function and using zeromodel and Zero Inflated Poisson [table 16] Genmod Procedure dist as ZINB, log as link function and using zeromodel for the nodes as response variable. The AIC of all four models were compared [table 17]. The AIC of Zero Inflated Negative Binomial was the least among all to predict the factors influencing the increase in the number of lymph nodes, as it accounts for both overdispersion, variability in the data and the number of zeros. The assumption of independence was checked using residual against observations using scatterplot procedure in SAS[table 18].

Results and Discussions

The Cox hazard proportional model was performed to do a survival analysis to find the influencing predictors that increase the mortality risk. The model was performed with stepwise selection which resulted in 6 significant variables suggesting that these siz variables plays a role in increasing the risk.

The summary of the Analysis [table7] is as follows:
Age: The hazard ratio for age is 1.022, indicating that for each additional year of age, the risk of mortality increases by a factor of 1.022, holding all other variables constant. The p-value for age is less than 0.0001, indicating that age is a significant predictor of the mortality.
Grade: The hazard ratio for grade is 1.253, indicating that for each increase in grade of the tumor, the risk of mortality by a factor of 1.253, holding all other variables constant. The p-value for grade is 0.0013, indicating that grade is a significant predictor of the event.
Nodes: The hazard ratio for nodes is 1.056, indicating that for each additional positive lymph node, the risk of mortality increases by a factor of 1.056, holding all other variables constant. The p-value for nodes is less than 0.0001, indicating that nodes is a significant predictor of the event.
PGR: The hazard ratio for PGR is 1.000, indicating that there is no significant effect of progesterone receptor levels on the risk of mortality.
ER: The hazard ratio for ER is 1.000, indicating that there is no significant effect of estrogen receptor levels on the risk of mortality. However, the p-value for ER is 0.0877, which is greater than the significance level of 0.05, indicating that the effect of ER is not statistically significant.
Recur: The hazard ratio for recur is 6.908, indicates that patients who experienced recurrence have a much higher risk of experiencing an event (relapse) compared to those who did not experience recurrence. The p-value for recur is less than 0.0001, indicating that recur is a highly significant predictor of the event.

The output suggests that, the predictors grade and recur are the only two factors that plays a crucial role in increasing the risk of mortality. For age, nodes, pgr and er, the hazard ratio is close to one having no significance in the risk of mortality. But the factor grade has an influence in the risk and the variable recurrence states that there is 6.9 times higher chance of mortality.

The survival plot for patients with breast cancer, stratified by whether or not they had a recurrence of breast cancer shows two curves, one for patients who had a recurrence of breast cancer (the bottom curve) and one for patients who did not have a recurrence (the top curve). The x-axis represents time in months, and the y-axis represents the proportion of patients who have not experienced the event of interest (in this case, death) at that time point. The plot suggests that patients who did not have a recurrence have a higher probability of survival compared to those who did have a recurrence. At the beginning of the study, the survival probability is close to 1.0 for both groups. However, over time, the survival probability decreases for both groups, indicating that patients are experiencing the event of interest (death). The decrease in survival probability is more rapid for patients who had a recurrence, as shown by the steeper slope of the curve. Overall, the plot suggests that recurrence of breast cancer is associated with worse survival outcomes in patients with breast cancer.

The survival curves for patients with breast cancer, stratified by whether or not they had a recurrence of breast cancer shows two curves, one for patients who had a recurrence of breast cancer (the bottom curve) and one for patients who did not have a recurrence (the top curve). The x-axis represents time in months, and the y-axis represents the proportion of patients who have not experienced the event of interest (in this case, death) at that time point. The plot suggests that patients who did not have a recurrence have a higher probability of survival compared to those who did have a

recurrence. At the beginning of the study, the survival probability is close to 1.0 for both groups. However, over time, the survival probability decreases for both groups, indicating that patients are experiencing the event of interest (death). The decrease in survival probability is more rapid for patients who had a recurrence, as shown by the steeper slope of the curve. Overall, the plot suggests that recurrence of breast cancer is associated with worse survival outcomes in patients with breast cancer.

The logistic regression model was fitted to predict the likelihood of death based on the following predictors: age, nodes, pgr, er, hormon, chemo, recur, dtime, grade, meno and recur. The analysis of maximum likelihood estimates shows that 6 predictors variables are significant predictors of death at the $p < .05$ level, as evidenced by their Wald chi-square statistics.

The most significant predictor variable is recur, with an estimate of 2.8631, which represents the log odds of death when recur is present. In comparison to no recurrence, the odds of death are 17.516 times higher when recurrence is present.

The summary of the table is as follows:
For a one-unit increase in age, the log odds of death increase by 1.041 holding all other variables constant.
For a one unit increase in age, the rate of death increase by 4.1%, holding all other variables constant.
For a one-unit increase in the number of positive lymph nodes, the log odds of death increase by 1.076, holding all other variables constant.
For a one unit increase in the number of positive lymph nodes, the rate of death increases by 7.6%, holding all other variables constant.
For a one-unit increase in the ER level, the log odds of death increase by 1, holding all other variables constant.
For a one unit increase in the ER level, the rate of death increase by 0%, holding all other variables constant.
For patients who have hormone therapy compared to those who don't, the log odds of death decrease by 0.317, holding all other variables constant.
For patients who have hormone therapy compared to those who don't, the rate of death decrease by 68.3%, holding all other variables constant.
For a one unit increase in the dtime, the rate of death decrease by 0.1%, holding all other variables constant.
For a one unit increase in the dtime, the log odds of death decrease by 0.999, holding all other variables constant.
For patients who had a recurrence of cancer compared to those who did not, the log odds of death increase by 2.8631, holding all other variables constant.
For patients who had a recurrence of cancer compared to those who did not, the rate of death increase by 1651.6% holding all other variables constant.
Looking at the output of Logistic Regression, it is evident that the factors age, lympnodes, hormon therapy and recurrence plays a significant role.

The factors Estrogen receptor and dtime doesn't make much of an influence in the rate of death.

Out of the 4 significant predictors, patient receiving hormon therapy has a lesser chance of death compared to patients who don't, which adds a meaningful value in the prognosis.

Considering the variables age, it is evident that as age increases mortality rate increases as well, also the nodes cause increase in the rate of death is reasonable. As the number of lymph nodes affected by cancerous cells increases, the number of tumor cells also grows which can potentially lead to the increase in the death rate. Finally, if recurrence of breast cancer occurs in patients already affected then it results in increase in the death rate, which also correlates with the findings from Cox Proportional model.

Performing the Zero Inflated Negative binomial with nodes as response variable, the result showed Pearson Chi sq Value/Df close to 1 indicating the model to be a good fit. Summary of the estimates are as follows:

Out of the 10 predictors used, only 6 were influential with p value lesser than 0.05, Summary of the estimates are as follows:

Interpretation of the coefficient using exponential function (exp):
grade: For a one unit increase in grade, the average number of lymph nodes increases by a factor of exp(0.1473) = 1.16, holding all other variables constant.

For a one unit increase in grade, the average number of lymph nodes increases by 15.8%, holding all other variables constant.
pgr: For a one unit increase in pgr, the average number of lymph nodes decreases by a factor of exp(-0.0002) = 0.9998, holding all other variables constant.
For a one unit increase in pgr, the average number of lymph nodes decreases by 0.02%, holding all other variables constant.
hormon: For a one unit increase in hormon, the average number of lymph nodes increases by a factor of exp(0.1259) = 1.13, holding all other variables constant.
For a one unit increase in hormon, the average number of lymph nodes increases by 13%, holding all other variables constant.
chemo: For a one unit increase in chemo, the average number of lymph nodes decreases by a factor of exp(-0.1610) = 0.85, holding all other variables constant.
For a one unit increase in chemo, the average number of lymph nodes decreases by 15%, holding all other variables constant.
rtime: For a one unit increase in rtime, the average number of lymph nodes decreases by a factor of exp(-0.0001) = 0.9999, holding all other variables constant.
For a one unit increase in rtime, the average number of lymph nodes decreases by 0.01%, holding all other variables constant.
recur: For a one unit increase in recur, the average number of lymph nodes increases by a factor of exp(0.2450) = 1.28, holding all other variables constant.
For a one unit increase in recur, the average number of lymph nodes increases by 28%, holding all other variables constant.
Here the pgr and rtime variables doesn't add any value to the interpretation as their rates are very less. Considering the pvalue and the output estimates, only 4 variables play a crucial role in increasing the number of nodes - grade, hormon therapy, chemo therapy and recurrence. Also here, performing hormon therapy shows increase in the lymph nodes but in real time, hormon therapy should decrease the affected lymph nodes, further analysis is required on this study of why hormon therapy is negatively impacting the number of nodes.

Conclusion

The study provides insights into the recurrence of breast cancer, the age distribution of affected patients, and the treatment options available, including hormone therapy and chemotherapy. The use of frequency distribution tables helps to understand the distribution of different variables in the dataset, which can be used to identify patterns and associations. The analysis of mortality and survival rates provides important information on the severity of the disease and its potential impact on patients' lives, allowing healthcare providers to better counsel patients and personalize treatment plans. Identifying predictors that influence time to death event in breast cancer patients using the Cox proportional hazards model helps the doctors personalize treatment plans and improve patient outcomes. Understanding which factors are associated with higher mortality rates can help identify high-risk patients who may require closer monitoring or more aggressive treatment. By analysing data from past breast cancer patients, Cox proportional hazards model can be used to inform clinical decision-making for future patients and potentially improve survival rates.

Identifying the predictors playing crucial role in the mortality rate of breast cancer patients using Logistic regression method, helps doctors to better assess the risks and benefits of different treatment options for each individual patient. Understanding which factors are associated with higher mortality rates can help healthcare providers to counsel patients on lifestyle changes or other interventions that may improve their chances of survival. Logistic regression on the rate of mortality can further help to identify subgroups of patients who are at higher risk for breast cancer mortality, which can be used to tailor screening programs and early interventions. This can be done with higher volume of data and more predictors.

Identifying predictors of the growth of tumor affected lymph nodes causing metastasis using the Zero-inflated negative binomial regression can help doctors better understand the reasons for the spread of cancerous cells which in turn helps doctor focus on the factors causing metastasis.

Identifying the factors associated with higher rates of lymph node metastasis, which can help healthcare providers to identify high-risk patients and potentially intervene earlier. By analysing data from past breast cancer patients, zero-inflated negative binomial regression can help inform clinical decision-making for future patients, potentially improving survival rates and quality of life.

Lastly all the three analysis suggest that patients already affected with breast cancer should be extra precautious as the recurrence of breast cancer can lead to adverse effects.

Limitations

In this dataset, the variable dtime was used in the Cox Proportional Hazard Model for predicting factors influencing the time till mortality. But there is chances for this variable to be having date until last follow up rather than date until death, since the information was not clear, I have taken the variable to be the time till death and continued with the modelling.

There can be different factors associated with predicting the prognosis of breast cancer as well as for the survival rate prediction such as hormon receptor poisitve or negative breast cancer, list of drugs taken by patients thus it is important to consider all those clinical factors when expecting for a higher accuracy on the breast cancer prognosis.

Also, this dataset consists of 2982 patient observation, to predict a disease such as breast cancer a huge chunk of data is required. This study is just a sample for what can be done even when there is higher volume of data.

References

Dataset from:
https://stat.ethz.ch/R-manual/R-devel/library/

Research reference:
Royston, P., & Altman, D. G. (2013). External validation of a Cox prognostic model: principles and methods. BMC medical research methodology, 13(1), 33.

Reference for SAS procedures:
https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/pgmsaswlcm/home.htm

Appendix

Code Book

Code 1: Data Load, Labelling and finding the missing values

```
LIBNAME glm '/home/u62518985';
FILENAME bcdata '/home/u62518985/GLM/breastcancer.csv';
proc import
   datafile = bcdata
   out = glm.breastcancerdata
   dbms = csv
   replace;
   getnames = yes;
run;
data glm.breastcancerdata;
   set glm.breastcancerdata;
   if size = "<=20" then tumor_size = 2;
   else if size = "20-50" then tumor_size = 1;
   else if size = ">50" then tumor_size = 0;
run;
data glm.breastcancerdata;
   set glm.breastcancerdata;
   label
```

```
      id='patient id'
      pid='patient id within study'
      year='year the patient underwent surgery'
      age='age of patient during surgery'
      meno='menopausal status (0=premenopausal, 1=postmenopausal)'
      size='size of tumor'
      grade='grade of tumor'
      nodes='number of positive lymph nodes'
      pgr='progesterone receptors(fmol/l)'
      er='estrogen receptors (fmol/l)'
      hormon='hormonal therapy (0=not received, 1=received)'
      chemo='chemotherapy received or not (0=not received, 1=received)'
      recurence_free='recurrence-free survival time (months) or last follow-up'
      rtime ='days to relapse or last follow-up)'
      recur = 'whether or not the patient experienced recurrence(0= no relapse, 1= relapse)'
      dtime = 'days to relapse or last follow-up'
      death='whether the patient died during the follow-up period 0=alive, 1=death'
      tumor_size='size of tumor, <=20 is 1, 20-50 is 2 and >50 is 3';
run;
proc means data=glm.breastcancerdata nmiss;
var year age meno tumor_size grade nodes pgr er hormon chemo rtime recur dtime death;
run;
```

Code 2: Descriptive Statistics

```
proc means data=glm.breastcancerdata n mean median min max stddev skewness;
   var age  ;
run;
proc univariate data=glm.breastcancerdata noprint;
   var  age;
   histogram / normal;
run;
proc univariate data=glm.breastcancerdata ;
   var age;
   qqplot / normal;
run;
proc corr data=glm.breastcancerdata;
var dtime rtime;
run;
proc freq data=glm.breastcancerdata;
table recur death;
run;
proc freq data=glm.breastcancerdata;
table chemo death;
run;
proc freq data=glm.breastcancerdata;
table hormon death;
run;
```

Code 3: Cox Proportional Hazard Model And Survival Curve

```
/*survival analysis*/
proc phreg data=glm.breastcancerdata;
  model dtime*death(0) = age meno grade nodes pgr er hormon chemo recur/selection=stepwise
slentry=0.20 slstay=0.15 covb;
run;
proc lifetest data=glm.breastcancerdata notable;
 time dtime*death(0);
strata recur;
run;
```

```
proc lifetest data=glm.breastcancerdata notable;
 time dtime*death(0);
strata grade;
run;
```

Code 4: Logistic Regression

```
proc logistic data= glm.breastcancerdata descending;
   model death = age meno tumor_size grade nodes pgr er hormon chemo rtime recur
dtime/selection=stepwise ;
   output out=glm.logpred resdev=resdev dfbetas=dfbetas pred ;
run;
data glm.breastcancerdata;
set glm.breastcancerdata;
observation=_n_;
run;
proc sgplot data=glm.logpred;
   scatter x=observation y=resdev / markerattrs=(symbol=circlefilled) markerfillattrs=(color=red);
   xaxis label="observation";
   yaxis label="residuals";
run;
```

Code 5: Poisson, Negative Binomal, Zero Inflated Poisson, Zero Inflated Negative Binomial And The Assumption Of Independence For Zinb

```
title 'number of zeros in nodes';
proc sql;
    select count(*) as num_zeros
    from glm.breastcancerdata
    where nodes = 0;
quit;
proc genmod data= glm.breastcancerdata;
   model nodes = age meno death grade pgr er hormon chemo rtime recur/ dist =poisson link=log;
run;
proc means data=glm.breastcancerdata mean var;
var nodes;
run;
proc genmod data=glm.breastcancerdata;
   model nodes = age meno death grade pgr er hormon chemo rtime recur/ dist =negbin link=log;
run;
proc genmod data=glm.breastcancerdata;
   model nodes = age meno death grade pgr er hormon chemo rtime recur/ dist=zip link=log;
   zeromodel age meno death pgr er hormon chemo rtime recur dtime ;
run;
proc genmod data=glm.breastcancerdata;
   model nodes = age meno death grade pgr er hormon chemo rtime recur/ dist=zinb link = log ;
   zeromodel age meno death pgr er hormon chemo rtime recur dtime  ;
   output out=glm.zinbresiduals pred=predicted reschi=reschi;
run;
data glm.breastcancerdata;
set glm.breastcancerdata;
observation=_n_;
run;
proc sgplot data=glm.zinbresiduals;
   scatter x=observation y=reschi / markerattrs=(symbol=circlefilled) markerfillattrs=(color=red);
   xaxis label="observation";
   yaxis label="residuals";
run;
```

Tables and Plots

Table 3: Summary of Nodes and Age

The MEANS Procedure

Analysis Variable : age Age of Patient during surgery						
N	Mean	Median	Minimum	Maximum	Std Dev	Skewness
2982	55.0583501	54.0000000	24.0000000	90.0000000	12.9529876	0.1103529

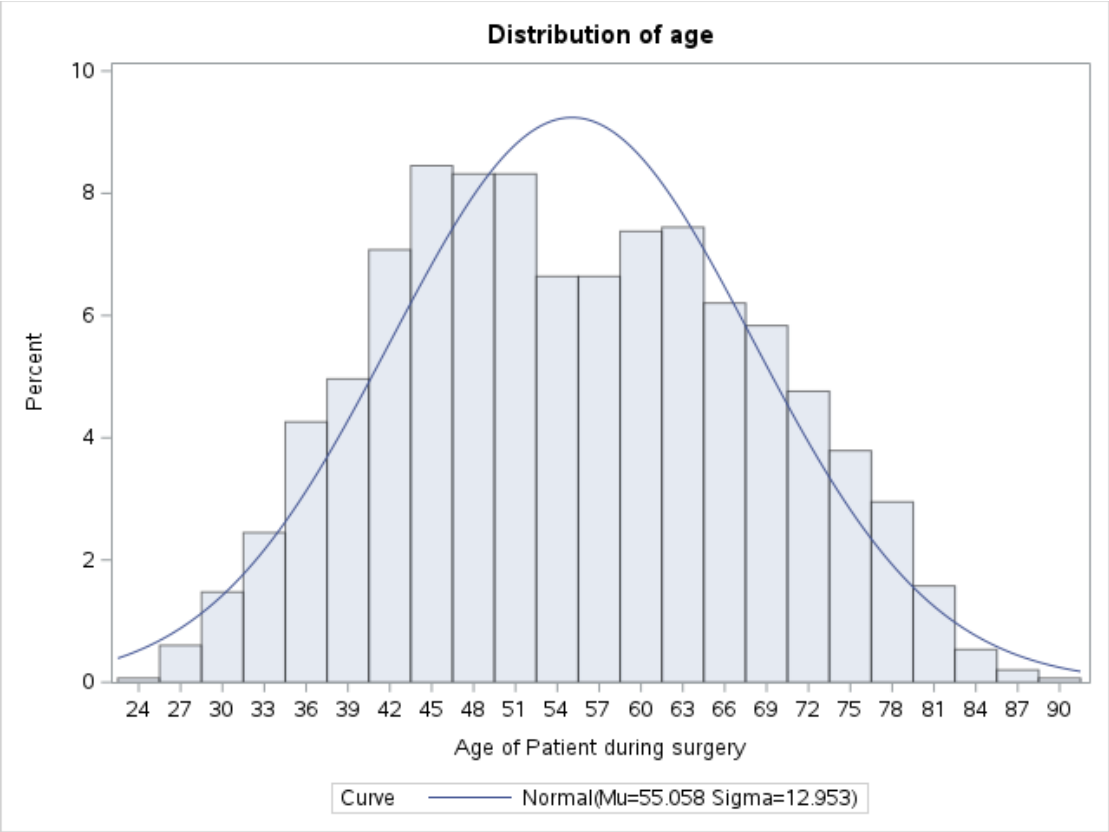


Table 4: Frequency distribution of Nodes Affected

The FREQ Procedure

Number of Positive Lymph Nodes				
nodes	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	1436	48.16	1436	48.16
1	367	12.31	1803	60.46
2	240	8.05	2043	68.51
3	157	5.26	2200	73.78
4	136	4.56	2336	78.34
5	108	3.62	2444	81.96
6	92	3.09	2536	85.04
7	66	2.21	2602	87.26
8	60	2.01	2662	89.27
9	53	1.78	2715	91.05
10	55	1.84	2770	92.89
11	38	1.27	2808	94.16
12	44	1.48	2852	95.64
13	23	0.77	2875	96.41
14	19	0.64	2894	97.05
15	27	0.91	2921	97.95
16	8	0.27	2929	98.22
17	15	0.50	2944	98.73
18	2	0.07	2946	98.79
19	7	0.23	2953	99.03
20	7	0.23	2960	99.26
21	3	0.10	2963	99.36
23	6	0.20	2969	99.56
24	5	0.17	2974	99.73
25	2	0.07	2976	99.80
27	2	0.07	2978	99.87
30	1	0.03	2979	99.90
34	3	0.10	2982	100.00

Table 5: Correlation between rtime and dtime

The CORR Procedure

2 Variables:	dtime rtime

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
dtime	2982	2605	1298	7769124	36.00000	7043	days to relapse or last follow-up
rtime	2982	2098	1398	6255949	36.00000	7043	days to relapse or last follow-up)

Pearson Correlation Coefficients, N = 2982
Prob > |r| under H0: Rho=0

	dtime	rtime
dtime days to relapse or last follow-up	1.00000	0.82834 <.0001
rtime days to relapse or last follow-up)	0.82834 <.0001	1.00000

Table 6: Frequency distribution for recurrence status, chemo and hormon therapy status against death

recurrence vs death

The FREQ Procedure

Whether or not the patient experienced recurrence(0= no relapse, 1= relapse)				
recur	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	1464	49.09	1464	49.09
1	1518	50.91	2982	100.00

Whether the patient died during the follow-up period 0=Alive, 1=Death				
death	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	1710	57.34	1710	57.34
1	1272	42.66	2982	100.00

chemo vs death

The FREQ Procedure

Chemotherapy received or not (0=Not Received, 1=Received)				
chemo	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	2402	80.55	2402	80.55
1	580	19.45	2982	100.00

Whether the patient died during the follow-up period 0=Alive, 1=Death				
death	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	1710	57.34	1710	57.34
1	1272	42.66	2982	100.00

hormon vs death

The FREQ Procedure

Hormonal Therapy (0=Not Received, 1=Received)				
hormon	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	2643	88.63	2643	88.63
1	339	11.37	2982	100.00

Whether the patient died during the follow-up period 0=Alive, 1=Death				
death	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	1710	57.34	1710	57.34
1	1272	42.66	2982	100.00

Table 7: Cox Proportional Hazard model on dtime

Step 8. Effect chemo is removed. The model contains the following effects:

age grade nodes pgr er recur

Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	19054.833	17806.577
AIC	19054.833	17818.577
SBC	19054.833	17849.467

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1248.2556	6	<.0001
Score	1302.6598	6	<.0001
Wald	983.2771	6	<.0001

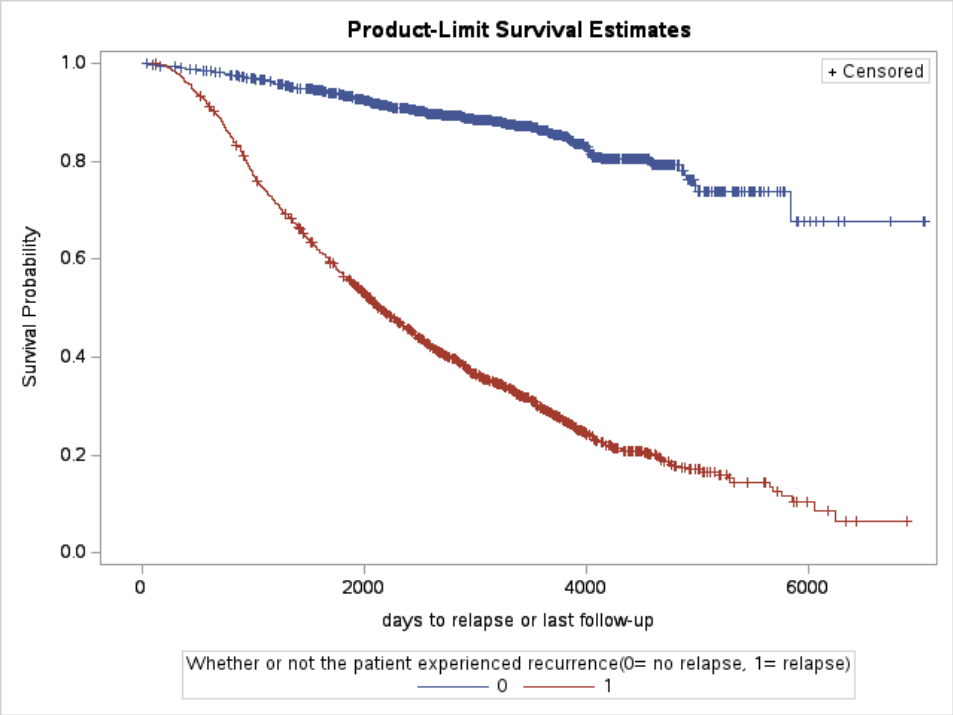
Note: Model building terminates because the effect to be entered is the effect that was removed in the last step.

Analysis of Maximum Likelihood Estimates							
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
age	1	0.02184	0.00237	84.7380	<.0001	1.022	Age of Patient during surgery
grade	1	0.22542	0.07028	10.2889	0.0013	1.253	Grade of Tumor
nodes	1	0.05450	0.00469	135.1918	<.0001	1.056	Number of Positive Lymph Nodes
pgr	1	-0.0004807	0.0001225	15.3950	<.0001	1.000	Progesterone Receptors(fmol/l)
er	1	-0.0001957	0.0001146	2.9154	0.0877	1.000	Estrogen Receptors (fmol/l)
recur	1	1.93267	0.07955	590.2424	<.0001	6.908	Whether or not the patient experienced recurrence(0= no relapse, 1= relapse)

Estimated Covariance Matrix							
Parameter		age	grade	nodes	pgr	er	recur
age	Age of Patient during surgery	0.0000056311	-.0000017158	-.0000007084	0.0000000148	-.0000000838	0.0000100229
grade	Grade of Tumor	-.0000017158	0.0049388297	-.0000072797	0.0000011222	-.0000002754	-.0003505380
nodes	Number of Positive Lymph Nodes	-.0000007084	-.0000072797	0.0000219740	0.0000000360	0.0000000124	-.0000629890
pgr	Progesterone Receptors(fmol/l)	0.0000000148	0.0000011222	0.0000000360	0.0000000150	-.0000000034	-.0000002172
er	Estrogen Receptors (fmol/l)	-.0000000838	-.0000002754	0.0000000124	-.0000000034	0.0000000131	-.0000003304
recur	Whether or not the patient experienced recurrence(0= no relapse, 1= relapse)	0.0000100229	-.0003505380	-.0000629890	-.0000002172	-.0000003304	0.0063282617

Summary of Stepwise Selection								
Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Effect Label
1	recur		1	1	945.7606		<.0001	Whether or not the patient experienced recurrence(0= no relapse, 1= relapse)
2	nodes		1	2	175.3384		<.0001	Number of Positive Lymph Nodes
3	age		1	3	81.5719		<.0001	Age of Patient during surgery
4	pgr		1	4	23.7353		<.0001	Progesterone Receptors(fmol/l)
5	grade		1	5	9.9596		0.0016	Grade of Tumor
6	er		1	6	2.9126		0.0879	Estrogen Receptors (fmol/l)
7	chemo		1	7	1.8808		0.1702	Chemotherapy received or not (0=Not Received, 1=Received)
8		chemo	1	6		1.8784	0.1705	Chemotherapy received or not (0=Not Received, 1=Received)

Table 8: Survival Curve using Recur as strata



Survival Curve using Grade as the strata

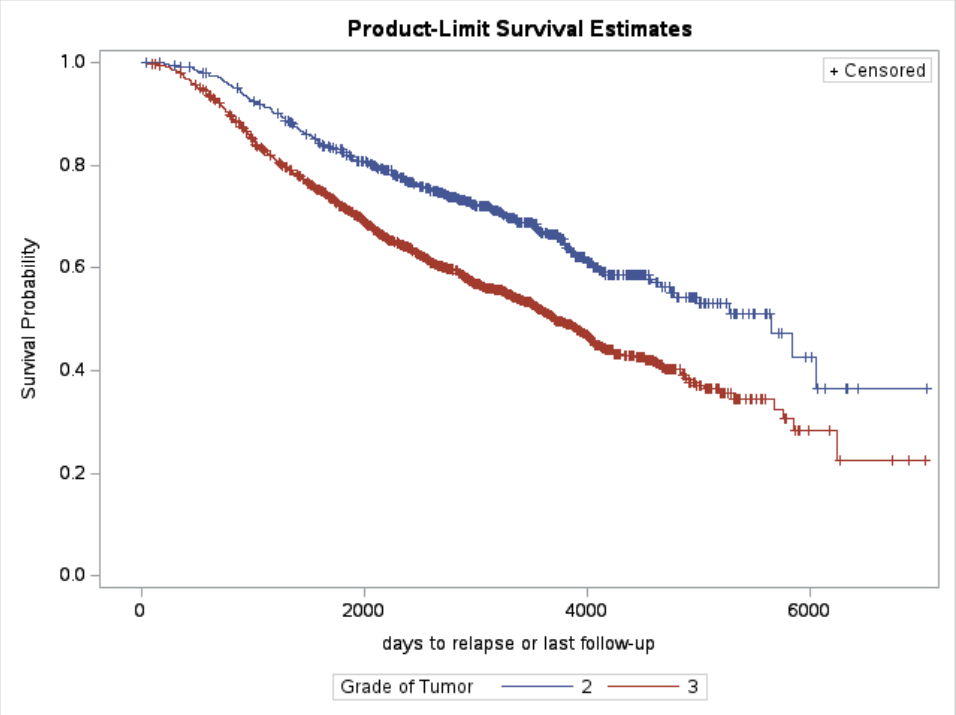


Table 9: Logistic Regression on death

The LOGISTIC Procedure

Model Information		
Data Set	GLM.BREASTCANCERDATA	
Response Variable	death	Whether the patient died during the follow-up period 0=Alive, 1=Death
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	2982
Number of Observations Used	2982

Response Profile		
Ordered Value	death	Total Frequency
1	1	1272
2	0	1710

Probability modeled is death='1'.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	4071.362	2144.987
SC	4077.363	2186.989
-2 Log L	4069.362	2130.987

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1938.3759	6	<.0001
Score	1531.2593	6	<.0001
Wald	861.0581	6	<.0001

Summary of Stepwise Selection								
Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Variable Label
1	recur		1	1	1011.8557		<.0001	Whether or not the patient experienced recurrence(0= no relapse, 1= relapse)
2	dtime		1	2	652.1506		<.0001	days to relapse or last follow-up
3	age		1	3	85.4974		<.0001	Age of Patient during surgery
4	hormon		1	4	28.7016		<.0001	Hormonal Therapy (0=Not Received, 1=Received)
5	nodes		1	5	24.1733		<.0001	Number of Positive Lymph Nodes
6	er		1	6	4.1653		0.0413	Estrogen Receptors (fmol/l)

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.4687	0.3087	22.6330	<.0001
age	1	0.0405	0.00473	73.2628	<.0001
nodes	1	0.0728	0.0148	24.3705	<.0001
er	1	0.000415	0.000203	4.1721	0.0411
hormon	1	-1.1503	0.1819	39.9984	<.0001
dtime	1	-0.00113	0.000054	429.0282	<.0001
recur	1	2.8631	0.1244	530.1183	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
age	1.041	1.032	1.051
nodes	1.076	1.045	1.107
er	1.000	1.000	1.001
hormon	0.317	0.222	0.452
dtime	0.999	0.999	0.999
recur	17.516	13.727	22.350

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	91.9	Somers' D	0.837
Percent Discordant	8.1	Gamma	0.837
Percent Tied	0.0	Tau-a	0.410
Pairs	2175120	c	0.919

Table 10: Assumption of Independence for Logistic Regression

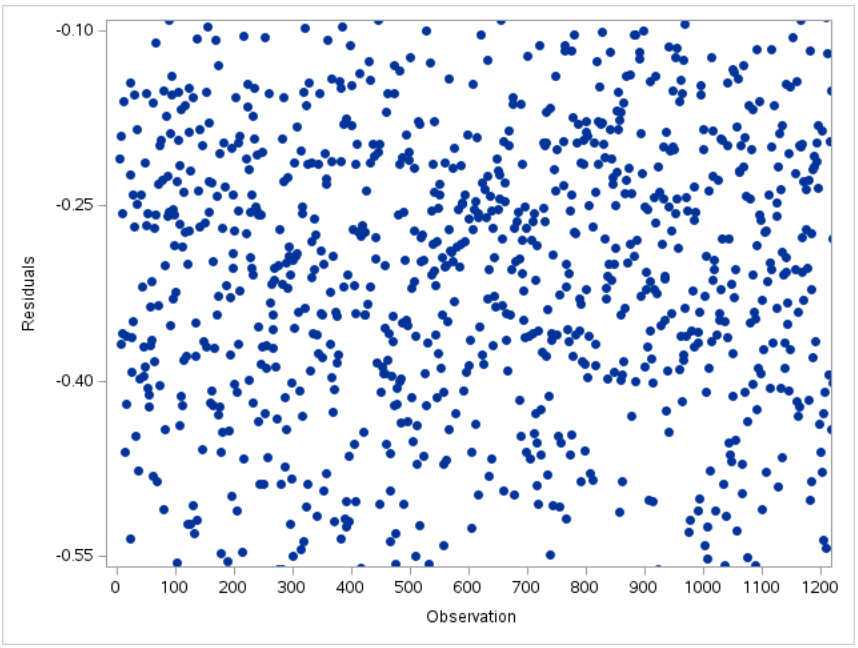


Table 11: Count of Zeroes in Nodes

Number of Zeros in Nodes

num_zeros
1436

Table 12: Poisson Regression

The GENMOD Procedure

Model Information		
Data Set	GLM.BREASTCANCERDATA	
Distribution	Poisson	
Link Function	Log	
Dependent Variable	nodes	Number of Positive Lymph Nodes

Number of Observations Read	2982
Number of Observations Used	2982

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	2971	11777.6062	3.9642
Scaled Deviance	2971	11777.6062	3.9642
Pearson Chi-Square	2971	16918.9548	5.6947
Scaled Pearson X2	2971	16918.9548	5.6947
Log Likelihood		2376.2468	
Full Log Likelihood		-8326.9826	
AIC (smaller is better)		16675.9652	
AICC (smaller is better)		16676.0541	
BIC (smaller is better)		16741.9691	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates						
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.6167	0.1167	-0.8454	-0.3880	27.92	<.0001
age	1	0.0120	0.0014	0.0092	0.0149	69.16	<.0001
meno	1	0.2760	0.0391	0.1994	0.3527	49.83	<.0001
grade	1	0.0928	0.0287	0.0365	0.1490	10.45	0.0012
death	1	0.3573	0.0335	0.2917	0.4229	113.88	<.0001
pgr	1	-0.0002	0.0000	-0.0003	-0.0001	10.75	0.0010
er	1	-0.0001	0.0000	-0.0002	0.0000	1.46	0.2270
hormon	1	0.7632	0.0278	0.7087	0.8178	751.92	<.0001
chemo	1	0.9612	0.0304	0.9016	1.0208	999.09	<.0001
recur	1	0.6108	0.0316	0.5489	0.6727	374.44	<.0001
dtime	1	-0.0002	0.0000	-0.0002	-0.0002	245.33	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

Note: The scale parameter was held fixed.

Table 13: Mean and Variance of Nodes

The MEANS Procedure

Analysis Variable : nodes Number of Positive Lymph Nodes	
Mean	Variance
2.7122736	19.2180915

Table 14: Negative Binomial Regression

The GENMOD Procedure

Model Information		
Data Set	GLM.BREASTCANCERDATA	
Distribution	Negative Binomial	
Link Function	Log	
Dependent Variable	nodes	Number of Positive Lymph Nodes

Number of Observations Read	2982
Number of Observations Used	2982

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	2971	2865.1727	0.9644
Scaled Deviance	2971	2865.1727	0.9644
Pearson Chi-Square	2971	4883.2506	1.6436
Scaled Pearson X2	2971	4883.2506	1.6436
Log Likelihood		5097.2584	
Full Log Likelihood		-5605.9710	
AIC (smaller is better)		11235.9420	
AICC (smaller is better)		11236.0471	
BIC (smaller is better)		11307.9462	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.3560	0.2853	-1.9151	-0.7969	22.60	<.0001
age	1	0.0135	0.0038	0.0060	0.0210	12.49	0.0004
meno	1	0.3780	0.1022	0.1778	0.5782	13.69	0.0002
grade	1	0.1468	0.0673	0.0149	0.2787	4.76	0.0292
death	1	0.5503	0.0837	0.3862	0.7144	43.19	<.0001
pgr	1	-0.0002	0.0001	-0.0004	0.0000	2.96	0.0851
er	1	0.0000	0.0001	-0.0002	0.0002	0.00	0.9675
hormon	1	1.0699	0.0887	0.8960	1.2437	145.50	<.0001
chemo	1	1.3650	0.0812	1.2058	1.5241	282.72	<.0001
recur	1	0.7247	0.0720	0.5836	0.8658	101.35	<.0001
dtime	1	-0.0001	0.0000	-0.0002	-0.0001	24.39	<.0001
Dispersion	1	1.7702	0.0740	1.6309	1.9213		

Note: The negative binomial dispersion parameter was estimated by maximum likelihood.

Table 15: Zero Inflated Poisson Regression

The GENMOD Procedure

Model Information		
Data Set	GLM.BREASTCANCERDATA	
Distribution	Zero Inflated Poisson	
Link Function	Log	
Dependent Variable	nodes	Number of Positive Lymph Nodes
Zero Model Link Function	Logit	

Number of Observations Read	2982
Number of Observations Used	2982

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance		11929.6126	
Scaled Deviance		11929.6126	
Pearson Chi-Square	2960	7558.2123	2.5535
Scaled Pearson X2	2960	7558.2123	2.5535
Log Likelihood		4738.4231	
Full Log Likelihood		-5964.8063	
AIC (smaller is better)		11973.6126	
AICC (smaller is better)		11973.9547	
BIC (smaller is better)		12105.6203	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	1.4824	0.1209	1.2455	1.7194	150.36	<.0001
age	1	-0.0015	0.0015	-0.0044	0.0015	0.96	0.3270
meno	1	-0.0086	0.0416	-0.0900	0.0729	0.04	0.8366
grade	1	0.1212	0.0290	0.0644	0.1780	17.48	<.0001
death	1	0.0662	0.0347	-0.0019	0.1342	3.64	0.0566
pgr	1	-0.0002	0.0000	-0.0003	-0.0001	19.34	<.0001
er	1	-0.0000	0.0000	-0.0001	0.0001	0.64	0.4231
hormon	1	0.0441	0.0287	-0.0122	0.1005	2.36	0.1244
chemo	1	-0.2510	0.0342	-0.3181	-0.1839	53.76	<.0001
recur	1	0.3550	0.0320	0.2923	0.4178	122.83	<.0001
dtime	1	-0.0001	0.0000	-0.0001	-0.0001	111.52	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

Note: The scale parameter was held fixed.

Analysis Of Maximum Likelihood Zero Inflation Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	3.7420	0.5221	2.7187	4.7653	51.37	<.0001
age	1	-0.0304	0.0071	-0.0444	-0.0165	18.21	<.0001
meno	1	-0.8444	0.1973	-1.2311	-0.4576	18.31	<.0001
grade	1	0.1075	0.1276	-0.1427	0.3577	0.71	0.3997
death	1	-1.1802	0.1421	-1.4587	-0.9018	69.00	<.0001
pgr	1	-0.0001	0.0002	-0.0005	0.0003	0.33	0.5637
er	1	0.0003	0.0002	-0.0001	0.0007	1.56	0.2111
hormon	1	-14.0631	40.7036	-93.8407	65.7145	0.12	0.7297
chemo	1	-15.3730	31.2336	-76.5898	45.8438	0.24	0.6226
recur	1	-0.9828	0.1311	-1.2398	-0.7258	56.19	<.0001
dtime	1	0.0001	0.0001	-0.0000	0.0002	1.98	0.1590

Table 16: Zero Inflated Negative Binomial Regression

The GENMOD Procedure

Model Information		
Data Set	GLM.BREASTCANCERDATA	
Distribution	Zero Inflated Negative Binomial	
Link Function	Log	
Dependent Variable	nodes	Number of Positive Lymph Nodes
Zero Model Link Function	Logit	

Number of Observations Read	2982
Number of Observations Used	2982

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance		9989.2708	
Scaled Deviance		9989.2708	
Pearson Chi-Square	2960	3790.8282	1.2807
Scaled Pearson X2	2960	3790.8282	1.2807
Log Likelihood		-4994.6354	
Full Log Likelihood		-4994.6354	
AIC (smaller is better)		10035.2708	
AICC (smaller is better)		10035.6440	
BIC (smaller is better)		10173.2788	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	1.3242	0.2375	0.8588	1.7897	31.10	<.0001
age	1	-0.0012	0.0030	-0.0071	0.0047	0.16	0.6905
meno	1	0.0048	0.0805	-0.1530	0.1626	0.00	0.9524
grade	1	0.1433	0.0542	0.0369	0.2496	6.97	0.0083
death	1	0.0463	0.0667	-0.0844	0.1771	0.48	0.4874
pgr	1	-0.0002	0.0001	-0.0004	-0.0000	6.18	0.0129
er	1	-0.0001	0.0001	-0.0003	0.0001	0.77	0.3798
hormon	1	0.1138	0.0584	-0.0007	0.2283	3.79	0.0515
chemo	1	-0.1793	0.0647	-0.3061	-0.0525	7.68	0.0056
recur	1	0.3959	0.0597	0.2790	0.5129	44.03	<.0001
dtime	1	-0.0001	0.0000	-0.0002	-0.0001	35.30	<.0001
Dispersion	1	0.5201	0.0287	0.4667	0.5796		

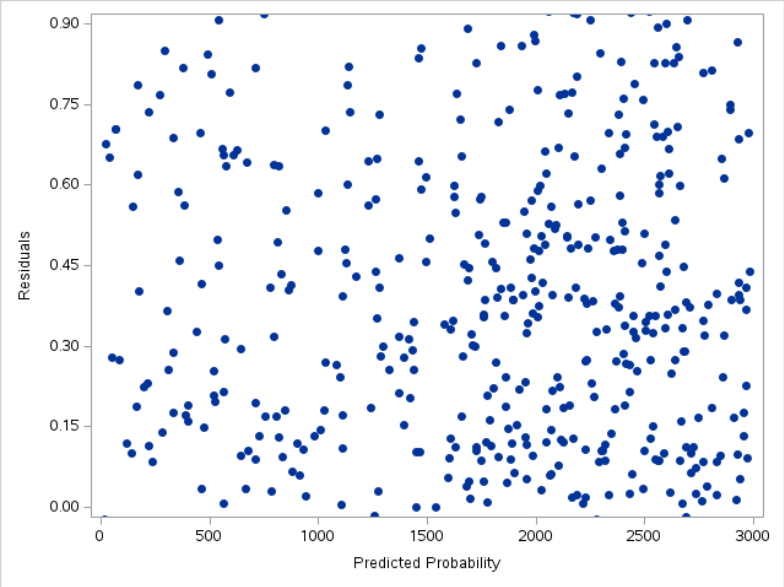
Note: The negative binomial dispersion parameter was estimated by maximum likelihood.

Analysis Of Maximum Likelihood Zero Inflation Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	3.7767	0.5728	2.6541	4.8994	43.47	<.0001
age	1	-0.0335	0.0079	-0.0490	-0.0180	18.03	<.0001
meno	1	-0.8582	0.2117	-1.2731	-0.4434	16.44	<.0001
grade	1	0.1459	0.1397	-0.1278	0.4197	1.09	0.2961
death	1	-1.2689	0.1573	-1.5772	-0.9605	65.05	<.0001
pgr	1	-0.0002	0.0002	-0.0006	0.0003	0.48	0.4888
er	1	0.0002	0.0002	-0.0002	0.0007	1.08	0.2994
hormon	1	-26.7573	24898.48	-48826.9	48773.37	0.00	0.9991
chemo	1	-28.1323	19350.90	-37955.2	37898.93	0.00	0.9988
recur	1	-0.9632	0.1432	-1.2439	-0.6826	45.24	<.0001
dtime	1	0.0000	0.0001	-0.0001	0.0002	0.58	0.4449

Table 17: AIC comparison

Poisson	**AIC (smaller is better)**	16672.22
Negative Binomial	**AIC (smaller is better)**	11235.88
Zero Inflated Poisson	**AIC (smaller is better)**	11973.18
Zero Inflated Negative Binomial	**AIC (smaller is better)**	10037.63

Table 18: Assumption of Independence for ZINB



----------END OF REPORT----------