**Survival Analysis based on Respiratory Cancer in Atlanta**

**Zheran Wang**

**Department of Mathematics and Statistics, Auburn University**

**12/5/2018**

# Abstract:

This analysis is analyzing the survival rate with month as response, and 9 other variables. In this study, 3 method are been used: Non-parametric Kaplan-Meier Estimator, Parametric regression model, and semi-parametric regression model. The conclusion of the study is the semi-parametric regression model is more suitable to the dataset.

# Introduction

## Background of SEER

SEER is a source provide by Surveillance Research Program (SRP) in NCI’s Division of Cancer Control and Population Sciences (DCCPS) for cancer statistics in the United States.

## Purpose of the Study

The purpose of this study is to analysis the respiratory cancer in Atlanta from year 1973 to 2015.

# Data

## Data source

SEER RESEARCHDATA RECORD DESCRIPTION

Website: <https://seer.cancer.gov>

“RESPIR” dataset has been selected amount all type of cancers.

## Variable Selection from Raw Data

In this study, 10 variables have been selected (Table 1). Amount them, survival month has been selected to be the response variable. SEER cause of death classification has been selected to be the status of cancer. Table 2 shows the description of each variables.

Table 3 shows the basic summary statistics about categorical variables, and table 4 shows the basic summary statistics about the continuous variables.

## Methods

The Methods of the analysis are listing below:

### Basic Survival plot

Using Proc lifetest in SAS to graph the survival plot and with different categorical groups.

### Parametric regression model

Using Proc lifereg in SAS to fit a parametric regression model.

Fit the model with Weibull, exponential, log normal, log logistic, and gamma distribution. Compare the AIC of each model.

Check model assumption

### Semi-Parametric regression model

Using Proc Phreg to fit a Cox Proportional Hazards Regression model with all parameter.

Check model adequacy by using Cox-Snell residual plot.

Check proportional Hazard assumption of each variable.

Using stepwise selection with pairwise interaction within continuous variable to select the model.

Change continuous variable nodes and extension into 3 different levels. Refit the model, and check adequacy.

Check if the dataset has outlier by using deviance residuals plot.

Check if stratified model is necessary.

# Result

## 3.1 Basic Survival plot.

Base on the overall survival plot in Figure 1. The plot shows that survival rate has a deep drop about 30 months after first diagnosis of cancer. Then the survival rate remains about the same.

Base on the survival plot in Figure 2. The plot shows that male tends to have lower survival rate compare to female. The difference between those two groups are not really large.

Base on the survival plot in Figure 3. This plot shows that there are not much difference in survival rate between single and marriage.

Base on the survival plot in Figure 4. This plot shows that black people tends to have lower survival rate than white. Other type of race has no differences between black and white. For further analysis. Race white will be treated as reference level in order to show significance.

Base on the survival plot in Figure 5. The plot shows that there is a clear difference between no primary side and left/right side. There is no difference between left side and right side. The patients have no primary site of the cancer tend to have higher survival rate compare to other type.

## 3.2 Parametric regression model.

Table 5 shows the fit statistics between 5 difference distribution assumption. The lowest AIC model was choosing to be the best model amount all 5 model, in this case, gamma distribution gives the lowest AIC value. Then use the model with to fit the parametric regression model with gamma as the distribution. The result of maximum likelihood estimator for parameters shows in table 6. Table 7 shows that all the groups have significant difference within each level. The marriage status result shows conflict with the survival plot. This due to the result after fitted model is by conditional calculation, which gives more accurate result compare to the survival plot.

Figure 6 shows the model assumption by fitting the model with gamma distribution. The SAS code is ProbPlot. If the distribution satisfied, most of the points need to be along a straight line. In this case, the points are not along a straight line. Therefore, this dataset is not well fitted by using parametric regression model.

## 3.3 Semi-Parametric regression model.

This time, a Cox Proportion Hazard regression model has been fitted to this data set. Table 8 shows the maximum likelihood estimator for parameters result with all the variables. Table 9 shows the group effect. In this case, the continuous variables significant shows the similar result with the parametric regression model. The marriage in the group effect shows insignificance, which is consistent with what has been observed by using survival plot.

After fitting the model, the model adequacy has been checked by using Cox-Snell residual plot (Figure 7). The plot shows after 2 in horizontal direction, the residual plot going away from the diagonal line. This means the model assumption doesn’t not hold. Transformation needs to apply to some of the variables.

The proportional hazards assumption has been checked on all continuous variables. Table 10 shows the result of the assumption test. The table shows that age, nodes, and extension variables need to apply some form of transformation to satisfy the assumption. Log transformation has applied to age, nodes, and extension. Figure 8, 9, and 10 shows the assumption check after the transformation. Base on the plot and the p-value shows in the plot, only log transformation on age is satisfied the model assumption. Therefore, other types of transformation need to apply onto nodes and extension.

Before apply transformation onto nodes and extension. A stepwise selection has been applied to the model, with pairwise interaction between continuous variables (Table 11). The model shown below:



This study ends up dividing nodes and extension into 3 categorical groups. For node, 0 to 5 nodes as level 0, 5 to 20 nodes as level 1, and nodes greater than 20 as level 2. For extension, 0 to 30 as level 0, 30 to 60 as level 1, and 60 to 99 as level 2. Then a new survival curve has been plotted base on the two new categorical group. Figure 11 shows the survival plot between different level of extension. This result shows with further extension of the tumor, the patients tend to have lower survival rate. Figure 12 shows the survival plot between different level of nodes. This result shows that with more regional lymph nodes been removed tends to have high survival rate. This result is consistent to clinical observations due to the lymph nodes system. Lymph nodes are the first organ the against tumor, which will be the easiest system that effected by cancer. Since lymph fluid will cycle within the whole body, it increases the change for the cancer to extent. Therefore, will more regional lymph nodes been removed have higher chance to survive.

Table 12 shows the model fitted with node and extension as categorical variables. Figure shows the overall model assumption appears better than the assumption before the transformation.

Figure 14 shows the deviance residual. The low absolute value shows no outlier in this dataset.

Figure 15, 16, 17, and 18 shows that treating certain categorical variable as strata to see if the model assumption will be better with stratified model. The results show this dataset doesn’t need stratified model.

# Conclusion

## 4.1 Non-Parametric regression model

Marriage status are not influence the cancer.

No primary site has significant different to left side and right side.

White and Black are significantly different to each other but has no difference to other type of race.

Different levels of extension have significantly different to each other.

More regional lymph nodes removed has higher survival rate.

## 4.2 Parametric regression model

Gamma distribution give lowest AIC value.

Model assumption seems not hold.

## 4.3 Semi-Parametric regression model.

Proportional hazards assumption seem hold.

Log transformation on age.

Change nodes and extension into categorical variables seem help.

No outlier.

No need for stratified model.

# Figures, Tables, and Graphs:

A close up of a device

Description automatically generated

# Table 1: Variable list

|  |  |  |
| --- | --- | --- |
| 1 | Month | Survival month(response) |
| 2 | Status | death (1), Censoring(death not causing by cancer) (0) |
| 3 | Birth | Year of Birth |
| 4 | Gender | Male (1), Female (0) |
| 5 | Marriage | Married (1), single (0) |
| 6 | Race | White(1), Black(2), Other(0) |
| 7 | Side | Side\_0=no side, side\_l=left side, side\_r=right side, side\_m=mideline |
| 8 | Agediag | Age of diagnosis |
| 9 | Extension | Furthest location of the cancer expands. In millimeter |
| 10 | Nodes | Total number of regional lymph nodes were removed. |
| 11 | Size | Tumor size. In millimeter |

\*For side, the cleaned dataset doesn’t have midline.

\*Highlight yellow means treated as baseline in the analysis.

# Table 2: Variable description

|  |  |  |  |
| --- | --- | --- | --- |
| Variable Name | Level | Frequency | Percent |
| Side | 0 | 302 | 4.98 |
|  | 1 | 3341 | 55.06 |
|  | 0 | 2425 | 39.96 |
| Gender | 0 | 1812 | 29.86 |
|  | 1 | 4256 | 70.14 |
| Race | 0 | 95 | 1.57 |
|  | 1 | 4425 | 72.92 |
|  | 2 | 1548 | 25.51 |
| Marriage | 1 | 1066 | 17.57 |
|  | 2 | 5002 | 82.43 |

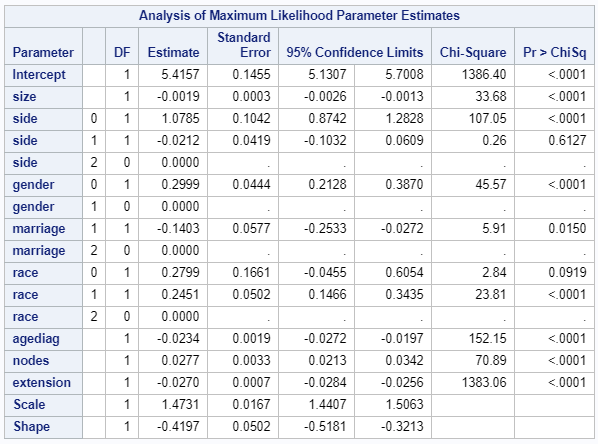
# Table 3: Statistic Summary of Categorical Variables



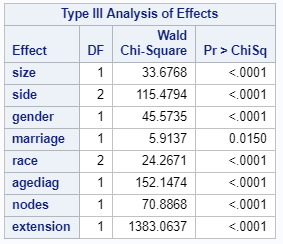
# Table 4: Statistic Summary of Continuous Variables

|  |  |  |  |
| --- | --- | --- | --- |
|  | -2 Log Likelihood | AIC | BIC |
| Weibull | 180375.56 | 18059.56 | 18139.17 |
| Exponential | 18770.58 | 18792.58 | 18865.56 |
| gamma | 17139.19 | 17165.19 | 17251.44 |
| Lognormal | 17211.32 | 17235.32 | 17314.93 |
| Log logistic | 17246.16 | 17270.16 | 17349.77 |

# Table 5: Fitted Statistics between 5 different distributions



# Table 6: Parametric Model with Gamma Distribution



# Table 7: Parametric Model with Group Effects



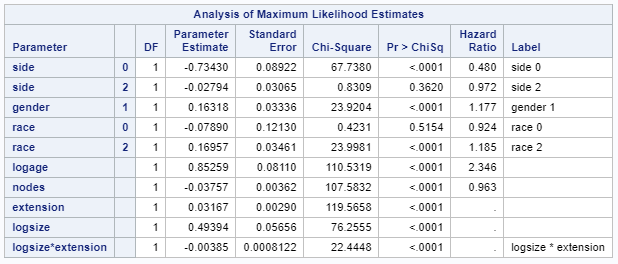
# Table 8: Semi-Parametric Model



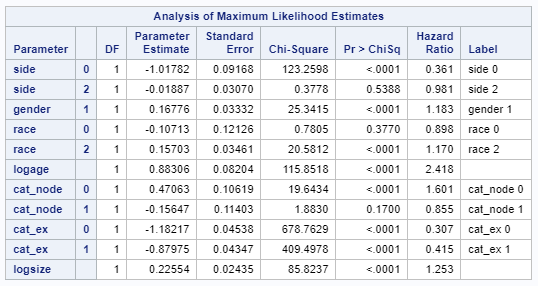
# Table 9: Type 3 Group Effect



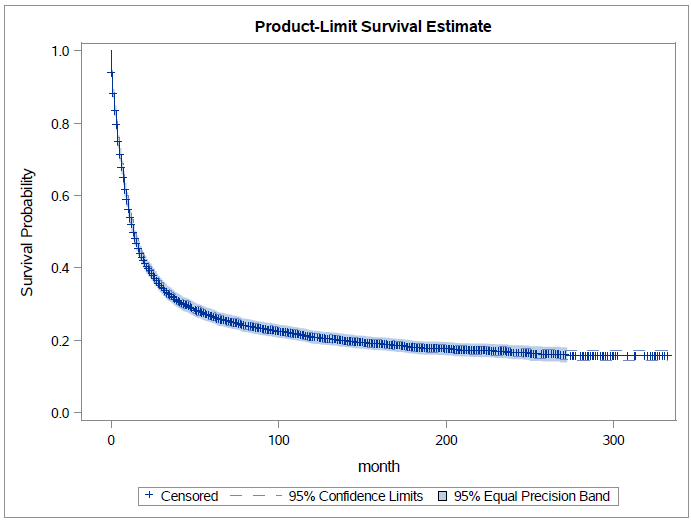
# Table 10: Supremum Test for Proportional Hazard Assumption



# Table 11: Stepwise Selection Result



# Table 12: Model Fitted with New Categorical Variables



# Figure 1: Overall survival plot

# 

# Figure 2: Survival plot with different gender



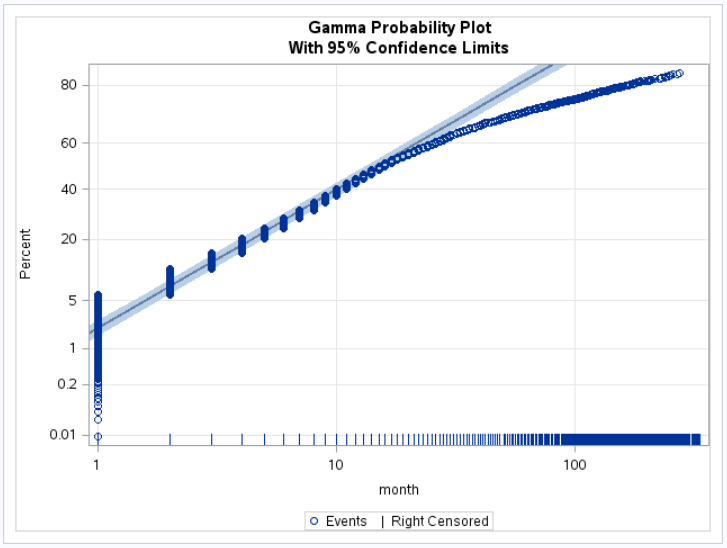
# Figure 3: Survival plot with Marriage Status



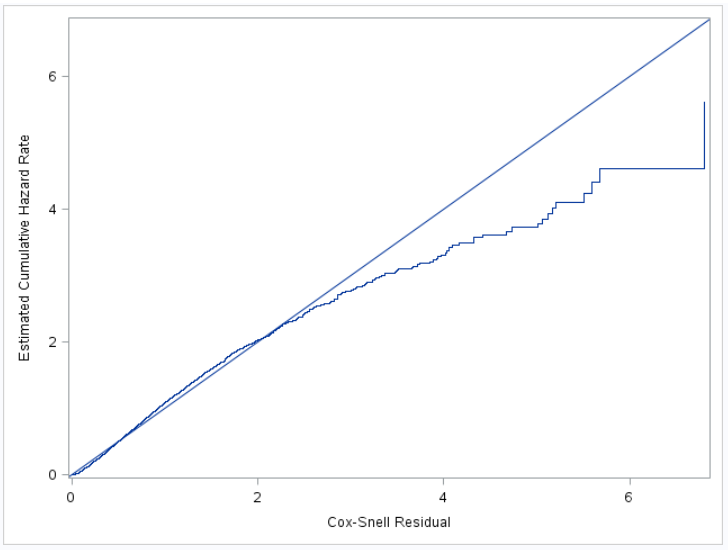
# Figure 4: Survival plot with different race



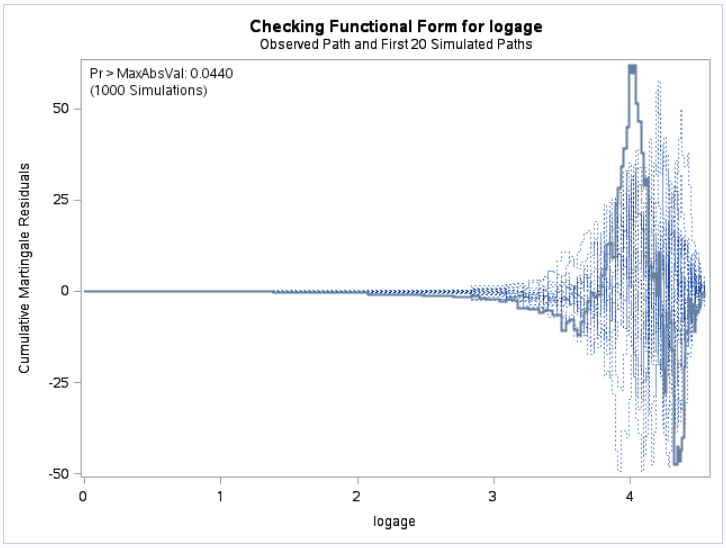
# Figure 5: Survival plot with different side



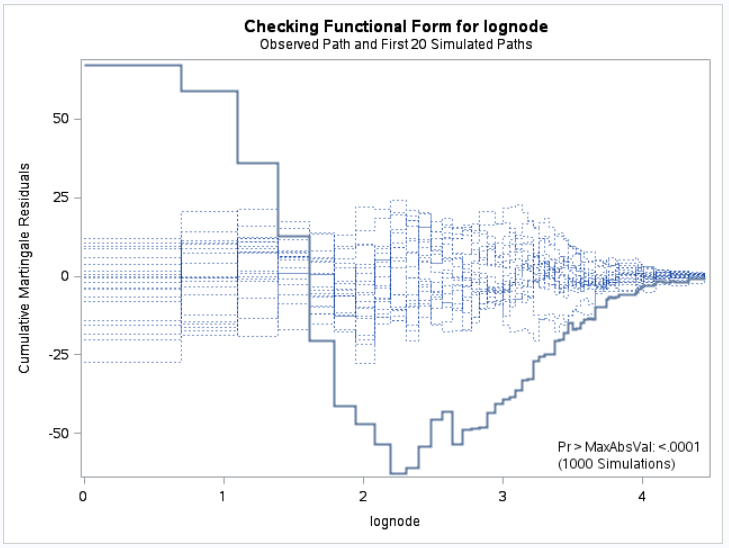
# Figure 6: Model Assumption of Gamma Distribution



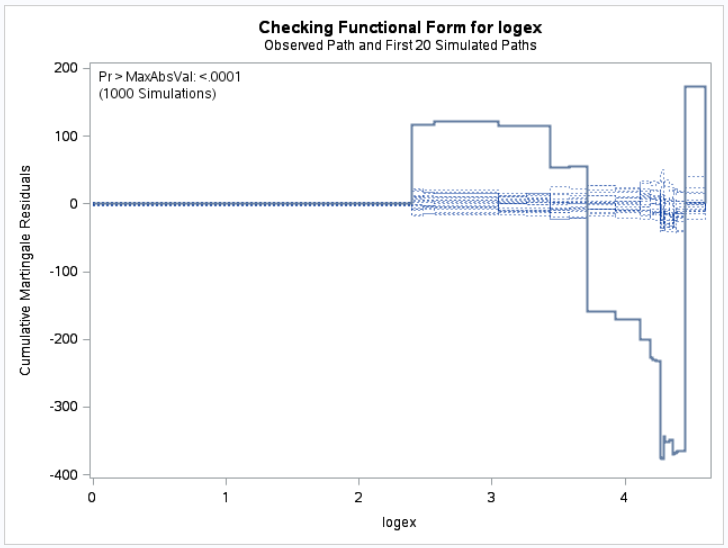
# Figure 7: Cox-Snell Residual Plot



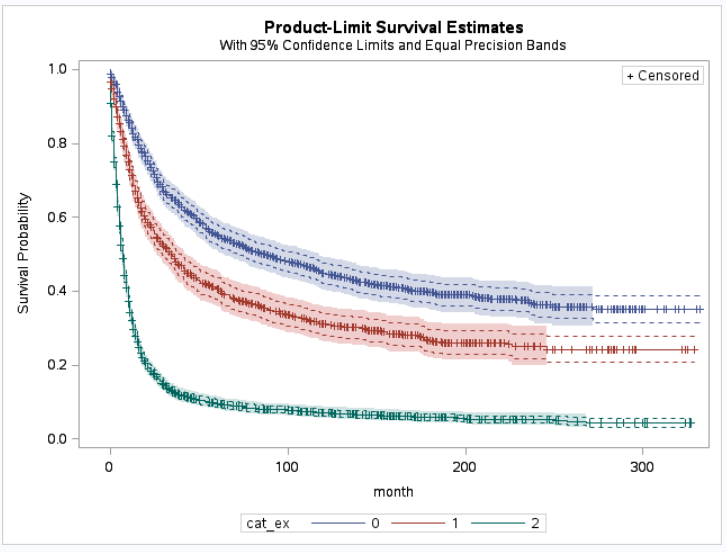
# Figure 8: Assumption Check of Log Age.



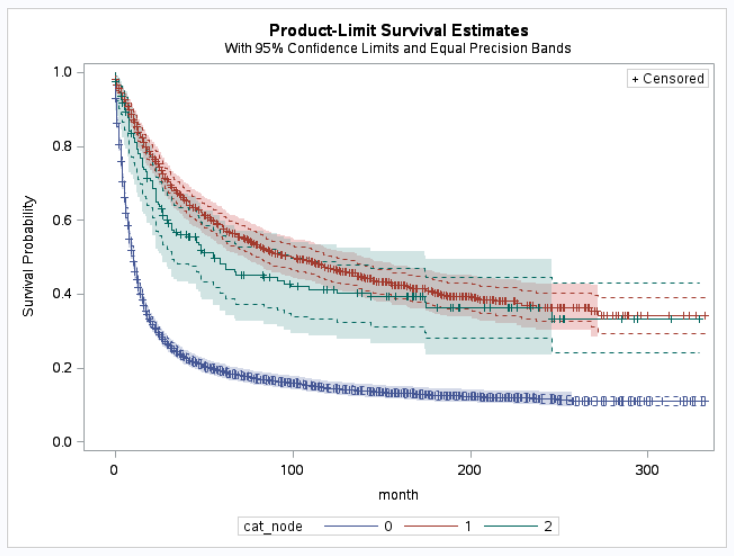
# Figure 9: Assumption Check of Log Nodes



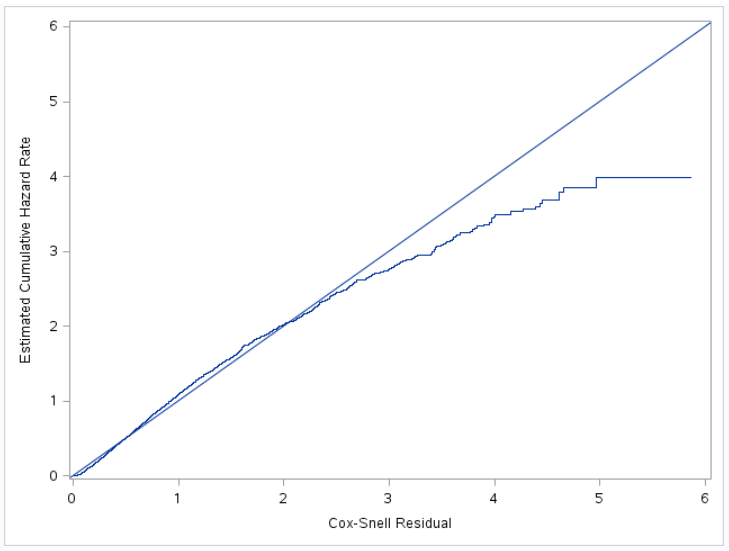
# Figure 10: Assumption Check of Log Extension



# Figure 11: Survival Plot between different levels of extension



# Figure 12: Survival Plot between different levels of nodes

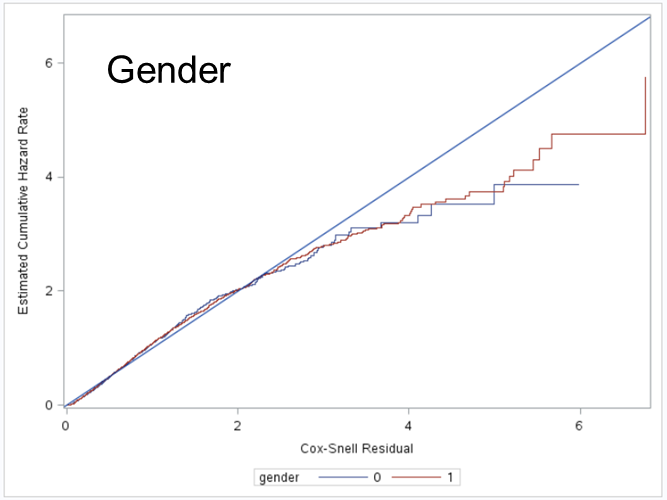


# Figure 13: Cox-Snell Residual Plot with New Model

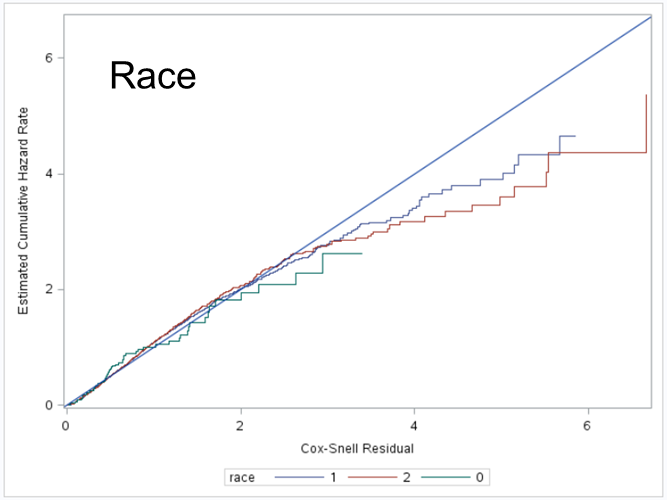
A screenshot of a cell phone

Description automatically generated

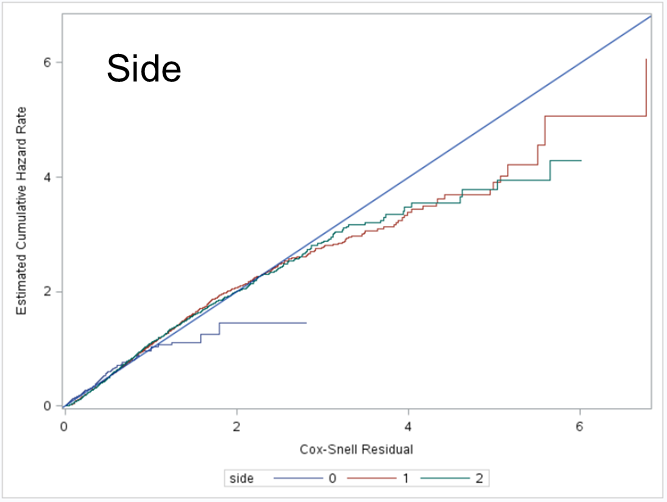
# Figure 14: Deviance Residual



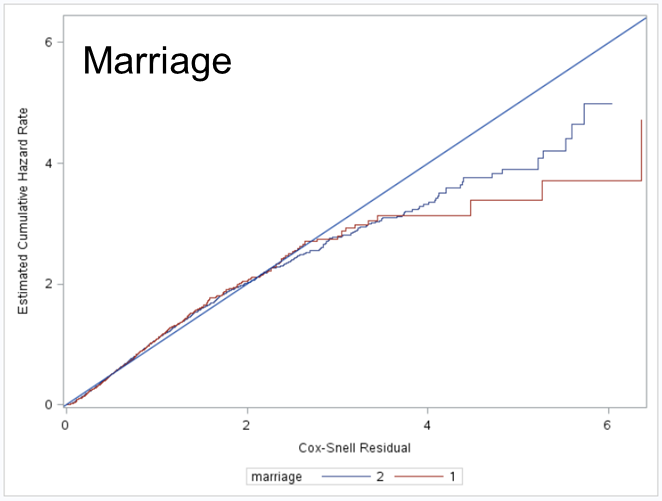
# Figure 15: Cox-Snell Residual with Gender Strata



# Figure 15: Cox-Snell Residual with Race Strata



# Figure 16: Cox-Snell Residual with Side Strata



# Figure 17: Cox-Snell Residual with Marriage Strata

# SAS CODE:

filename seer9 'I:\project\7780\RESPIR.TXT';

data in;

infile seer9 lrecl=362;

input

@1 PUBCSNUM $char8./\*patient ID\*/

@9 REG $char10./\*SEER registry\*/

@301 SRV\_TIME\_MON $char4./\*survival months\*/

@19 MAR\_STAT $char1./\*marital Status at DX \*/

@28 YR\_BRTH $char4./\*year of birth\*/

@25 AGE\_DX $char3./\*Age at Diagnosis\*/

@24 SEX $char1./\*gender of patient\*/

@20 RACE1V $char2./\*race of patient\*/

@47 LATERAL $char1./\*the initial side of tumor\*/

@61 EOD10\_SZ $char3./\*EOD-tumer size\*/

@64 EOD10\_EX $char2./\*EOD-extension\*/

@71 EOD10\_NE $char2./\*number of regional nodes examined\*/

@272 VSRTSADX $char1./\*SEER cause of death classification\*/

;

proc contents data=in;run;

proc freq data=in;

table race1v;

run;

data clean;

set in;

month =input(SRV\_TIME\_MON,4.);

status =input(VSRTSADX,1.);

marriage =input(MAR\_STAT,1.);

birth =input(YR\_BRTH,4.);

agediag =input(AGE\_DX,3.);

side =input(LATERAL,1.);

sex =input(sex,1.);

size =input(EOD10\_SZ,3.);

extension =input(EOD10\_EX,2.);

nodes =input(EOD10\_NE,2.);

if REG ='0000001527'; /\*Atlanta\*/

if RACE1V='01' then race=1; /\*creat dummies for race\*/

else if RACE1V='02' then race=2; /\*three race group, white,balck,other\*/

else if RACE1V GT '02' then race=0;

if sex='1' then gender=1; else gender=0;/\*dummy for sex\*/

if cmiss(of \_all\_) then delete;/\*delete missing values\*/

keep month status marriage birth agediag side

gender race size extension nodes;

proc contents data=clean;run;

data clean2;

set clean;

if month<9999;/\*unknow with month greater than 9999\*/

if agediag<998;/\*known age\*/

if marriage<3;/\*only consider single and married\*/

if status<2;/\*live(death not because of cancer) or death\*/

if side=0 or side=1 or side=2 or side=5;

/\*no side, left, right, pair(midline)\*/

if size NE 999;/\*999 means unknown size\*//\*in \*/

if extension<100;/\*allow values 00-99\*/

if nodes<90;/\*only consider exact number of nodes\*/

proc contents data=clean2;run;

/\*basic analysis\*/

proc means data=clean2 n min max mean std;

var agediag size extension nodes;

run;

proc export data=clean2

dbms=csv outfile="I:\project\7780\cancer.csv" replace;

run;

/\* Generated Code (IMPORT) \*/

/\* Source File: cancer.csv \*/

/\* Source Path: /home/zzw00490/STAT7780/Project \*/

/\* Code generated on: 11/28/18, 10:42 AM \*/

/\*import the clean data\*/

FILENAME REFFILE '/home/zzw00490/STAT7780/Project/cancer.csv';

PROC IMPORT DATAFILE=REFFILE

DBMS=CSV replace

OUT=WORK.cancer;

GETNAMES=YES;

RUN;

PROC CONTENTS DATA=WORK.cancer; RUN;

/\*un necessary step\*/

data cancer1;

set cancer;

drop birth;/\*not really easy to find out the age, so just drop it\*/

proc CONTENTS DATA=WORK.cancer1; RUN;

/\*frequency table\*/

proc freq data=cancer1;

table race gender marriage side;

run;

proc means maxdec=3 data=cancer1;

var size side agediag nodes extension;

run;

/\*\*\*\*some graphes of response\*\*\*\*/

/\*histogram\*/

proc univariate data=cancer1;

var month;

histogram month /kernel;

ods select Histogram;

run;

/\*cdf\*/

proc univariate data = cancer1;

var month;

cdfplot month;

ods select CDFPlot;

run;

/\*\*\*\*life test\*\*\*\*/

/\*nonparametric Kaplan Meier estimator.\*/

proc lifetest data=cancer1

outsurv = cancer1\_fitted alpha = 0.05 plots = S(CL CB = EP);

time month\*status(0);

ods select CensoredSummary SurvivalPlot Means Quartiles;

run;

/\*\*\*\*with different group\*\*\*\*/

/\*gender\*/

proc sort data=cancer1;by gender;run;

proc lifetest data=cancer1

outsurv = cancer1\_fitted alpha = 0.05 plots = S(CL CB = EP);

strata gender;

time month\*status(0);

ods select SurvivalPlot HomTests;

title 'Survival Plot with Different Gender';

run;

/\*race\*/

proc lifetest data=cancer1

outsurv = cancer\_fitted alpha = 0.05 plots = S(CL CB = EP);

strata race;

time month\*status(0);

ods select SurvivalPlot HomTests;

title 'Survival Plot with Different race';

run;

/\*marriage\*/

proc lifetest data=cancer1

outsurv = cancer\_fitted alpha = 0.05 plots = S(CL CB = EP);

strata marriage;

time month\*status(0);

ods select SurvivalPlot HomTests;

title 'Survival Plot with Marriage';

run;

/\*side\*/

proc lifetest data=cancer1

outsurv = cancer\_fitted alpha = 0.05 plots = S(CL CB = EP);

strata side;

time month\*status(0);

ods select SurvivalPlot HomTests;

title 'Survival Plot with Different sides';

run;

/\*\*\*\*lifereg\*\*\*\*/

/\*\*\*\*with assumption\*\*\*\*/

proc lifereg data = cancer1;

class side gender marriage race;

model month\*status(0)= size side gender marriage race agediag nodes extension;/\*weibull-special of gamma\*/

probplot;

ods select FitStatistics ProbPlot;

run;

/\*exponental-special of weibull\*/

proc lifereg data = cancer1;

class side gender marriage race;

model month\*status(0)= size side gender marriage race agediag nodes extension/dist=exponential;

ods select FitStatistics ProbPlot;

probplot;

run;

/\*gamma\*//\*lowest\*/

proc lifereg data = cancer1;

class side gender marriage race;

model month\*status(0)= size side gender marriage race agediag nodes extension/dist=gamma;

probplot;

run;

/\*lognormal-special of gamma\*/

proc lifereg data = cancer1;

class side gender marriage race;

model month\*status(0)= size side gender marriage race agediag nodes extension/dist=lnormal;

ods select FitStatistics ProbPlot;

probplot;

run;

/\*loglogistic\*/

proc lifereg data = cancer1;

class side gender marriage race;

model month\*status(0)= size side gender marriage race agediag nodes extension/dist=llogistic;

ods select FitStatistics ProbPlot;

probplot;

run;

/\*\*\*\*simple phreg\*\*\*\*/

proc phreg data=cancer1;

class side(ref='1') gender(ref='0') marriage(ref='1') race(ref='1');

model month\*status(0)= size side gender marriage race agediag nodes extension;

assess ph/resample=100;

run;

/\*\*\*\*transformation on variable?\*\*\*\*/

/\*node\*/

proc phreg data=cancer1;

class side(ref='1') gender(ref='0') marriage(ref='1') race(ref='1');

model month\*status(0)= size side gender race agediag lognode extension;

logage=log(agediag+1);

lognode=log(nodes+1);

logex=log(extension+1);

assess var=(lognode)/resample;

run;

/\*age\*/

proc phreg data=cancer1;

class side(ref='1') gender(ref='0') marriage(ref='1') race(ref='1');

model month\*status(0)= size side gender race logage nodes extension;

logage=log(agediag+1);

lognode=log(nodes+1);

logex=log(extension+1);

assess var=(logage)/resample;

run;

/\*extension\*/

proc phreg data=cancer1;

class side(ref='1') gender(ref='0') marriage(ref='1') race(ref='1');

model month\*status(0)= size side gender race agediag nodes logex;

logage=log(agediag+1);

lognode=log(nodes+1);

logex=log(extension+1);

assess var=(logex)/resample;

run;

/\*size\*/

proc phreg data=cancer1;

class side(ref='1') gender(ref='0') marriage(ref='1') race(ref='1');

model month\*status(0)= logsize side gender race agediag nodes extension;

logage=log(agediag+1);

lognode=log(nodes+1);

logex=log(extension+1);

logsize=log(size);

assess var=(logsize)/resample;

run;

/\*size, age cannot equal to zero\*/

proc phreg data=cancer1;

class side(ref='1') gender(ref='0') marriage(ref='1') race(ref='1');

model month\*status(0)= size side gender race logage nodes extension;

logage=log(agediag);

lognode=log(nodes+1);

logex=log(extension+1);

assess var=(logage)/resample;

run;

/\*stepwise with interaction on size, extension with nodes\*/

proc phreg data=cancer1;

class side(ref='1') gender(ref='0') race(ref='1');

model month\*status(0)= side gender race logage nodes extension logsize logage\*nodes extension\*logsize nodes\*extension logsize\*logage

/selection=stepwise ;

logage=log(agediag);

logsize=log(size);

run;

/\*only want logsize\*extension interactiong\*/

/\*final model\*/

proc phreg data=cancer1;

class side(ref='1') gender(ref='0') race(ref='1');

model month\*status(0)= side gender race logage nodes extension logsize extension\*logsize;

logage=log(agediag);

logsize=log(size);

run;

/\*\*\*\*assumption\*\*\*\*/

proc phreg data=cancer1 noprint;

class side(ref='0') gender(ref='0') marriage(ref='1') race(ref='0');

model month\*status(0)= size side gender marriage race logage nodes extension;

output out = cancer1\_fitted logsurv = CoxSnell;

logage=log(agediag+1);

run;

/\*Cox-Snell residual\*/

data cancer1\_fitted;

set cancer1\_fitted;

CoxSnell = -CoxSnell;

label CoxSnell = 'Cox-Snell Residual';

run;

/\*Nelson-Aalen estimate\*/

proc phreg data = cancer1\_fitted noprint;

model CoxSnell\*status(0)= ;

output out = cancer1\_fitted2 logsurv = haz;

run;

data cancer1\_fitted2;

set cancer1\_fitted2;

haz = -haz;

label haz = 'Estimated Cumulative Hazard Rate';

run;

proc sort data = cancer1\_fitted2; by CoxSnell;

proc sgplot data = cancer1\_fitted2 noautolegend;

step x = CoxSnell y = haz;

lineparm x=0 y=0 slope=1;

run;/\*good enough\*/

/\*\*\*\*strata?\*\*\*\*/

/\*marriage\*/

proc phreg data=cancer1 noprint;

class side(ref='0') gender(ref='0') marriage(ref='1') race(ref='0');

model month\*status(0)= size side gender race logage nodes extension;

strata marriage;

logage=log(agediag+1);

output out = cancer1\_stratified logsurv = CoxSnell;

run;

data cancer1\_stratified;

set cancer1\_stratified;

CoxSnell = -CoxSnell;

label CoxSnell = 'Cox-Snell Residual';

run;

proc sort data = cancer1\_stratified; by marriage;

proc phreg data = cancer1\_stratified noprint;

model CoxSnell \* status(0) = ;

output out = cancer1\_stratified2 logsurv = haz / method = ch;

by marriage;

run;

data cancer1\_stratified2;

set cancer1\_stratified2;

haz = -haz;

label haz = 'Estimated Cumulative Hazard Rate';

run;

proc sort data = cancer1\_stratified2; by CoxSnell;

proc sgplot data = cancer1\_stratified2;

step x = CoxSnell y = haz / group = marriage;

lineparm x=0 y=0 slope=1;

xaxis label = 'Cox-Snell Residual';

yaxis label = 'Estimated Cumulative Hazard Rate';

run;

/\*gender\*/

proc phreg data=cancer1 noprint;

class side(ref='0') gender(ref='0') marriage(ref='1') race(ref='0');

model month\*status(0)= size side marriage race logage nodes extension;

strata gender;

logage=log(agediag+1);

output out = cancer1\_stratified logsurv = CoxSnell;

run;

data cancer1\_stratified;

set cancer1\_stratified;

CoxSnell = -CoxSnell;

label CoxSnell = 'Cox-Snell Residual';

run;

proc sort data = cancer1\_stratified; by gender;

proc phreg data = cancer1\_stratified noprint;

model CoxSnell \* status(0) = ;

output out = cancer1\_stratified2 logsurv = haz / method = ch;

by gender;

run;

data cancer1\_stratified2;

set cancer1\_stratified2;

haz = -haz;

label haz = 'Estimated Cumulative Hazard Rate';

run;

proc sort data = cancer1\_stratified2; by CoxSnell;

proc sgplot data = cancer1\_stratified2;

step x = CoxSnell y = haz / group = gender;

lineparm x=0 y=0 slope=1;

xaxis label = 'Cox-Snell Residual';

yaxis label = 'Estimated Cumulative Hazard Rate';

run;

/\*side\*/

proc phreg data=cancer1 noprint;

class side(ref='0') gender(ref='0') marriage(ref='1') race(ref='0');

model month\*status(0)= size race marriage gender logage nodes extension;

strata side;

logage=log(agediag+1);

output out = cancer1\_stratified logsurv = CoxSnell;

run;

data cancer1\_stratified;

set cancer1\_stratified;

CoxSnell = -CoxSnell;

label CoxSnell = 'Cox-Snell Residual';

run;

proc sort data = cancer1\_stratified; by side;

proc phreg data = cancer1\_stratified noprint;

model CoxSnell \* status(0) = ;

output out = cancer1\_stratified2 logsurv = haz / method = ch;

by side;

run;

data cancer1\_stratified2;

set cancer1\_stratified2;

haz = -haz;

label haz = 'Estimated Cumulative Hazard Rate';

run;

proc sort data = cancer1\_stratified2; by CoxSnell;

proc sgplot data = cancer1\_stratified2;

step x = CoxSnell y = haz / group = side;

lineparm x=0 y=0 slope=1;

xaxis label = 'Cox-Snell Residual';

yaxis label = 'Estimated Cumulative Hazard Rate';

run;

/\*race\*/

proc phreg data=cancer1 noprint;

class side(ref='0') gender(ref='0') marriage(ref='1') race(ref='0');

model month\*status(0)= size side marriage gender logage nodes extension;

strata race;

logage=log(agediag+1);

output out = cancer1\_stratified logsurv = CoxSnell;

run;

data cancer1\_stratified;

set cancer1\_stratified;

CoxSnell = -CoxSnell;

label CoxSnell = 'Cox-Snell Residual';

run;

proc sort data = cancer1\_stratified; by race;

proc phreg data = cancer1\_stratified noprint;

model CoxSnell \* status(0) = ;

output out = cancer1\_stratified2 logsurv = haz / method = ch;

by race;

run;

data cancer1\_stratified2;

set cancer1\_stratified2;

haz = -haz;

label haz = 'Estimated Cumulative Hazard Rate';

run;

proc sort data = cancer1\_stratified2; by CoxSnell;

proc sgplot data = cancer1\_stratified2;

step x = CoxSnell y = haz / group = race;

lineparm x=0 y=0 slope=1;

xaxis label = 'Cox-Snell Residual';

yaxis label = 'Estimated Cumulative Hazard Rate';

run;

/\*\*\*\*\* deviance residuals \*\*\*\*\*/

proc phreg data = cancer1;

class side(ref='0') gender(ref='0') marriage(ref='1') race(ref='0');

model month\*status(0)= size side marriage gender logage nodes extension;

logage=log(agediag+1);

output out = fitted1 resmart = mgale resdev = resdev xbeta = risk;

run;

proc sgplot data = fitted1;

scatter x = risk y = mgale;

run;

proc sgplot data = fitted1;

scatter x = risk y = resdev;

run;/\*doesn't have any outlier\*/

/\*\*\*\*\*\*\*\*\*change into catagorical\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

data cancer2;

set cancer1;

if nodes<'5' then cat\_node=0;

else if nodes GE '5' and nodes<'20' then cat\_node=1;

else if nodes GE '20' then cat\_node=2;

if extension<'30' then cat\_ex=0;

else if extension GE '30' and extension<'60' then cat\_ex=1;

else if extension GE '60' then cat\_ex=2;

proc contents data=cancer2;run;

proc freq data=cancer2;

table cat\_node cat\_ex;

run;

/\*phreg\*/

proc phreg data=cancer2;

class side(ref='1') gender(ref='0') race(ref='1') cat\_node cat\_ex;

model month\*status(0)= side gender race logage cat\_node cat\_ex logsize;

logage=log(agediag);

logsize=log(size);

run;

/\*new cat plot\*/

proc sort data=cancer1;by cat\_node;run;

proc lifetest data=cancer2

outsurv = cancer1\_fitted alpha = 0.05 plots = S(CL CB = EP);

strata cat\_node;

time month\*status(0);

ods select SurvivalPlot HomTests;

title 'Survival Plot with level of node';

run;

proc sort data=cancer1;by cat\_ex;run;

proc lifetest data=cancer2

outsurv = cancer1\_fitted alpha = 0.05 plots = S(CL CB = EP);

strata cat\_ex;

time month\*status(0);

ods select SurvivalPlot HomTests;

title 'Survival Plot with level of extension';

run;

/\*\*\*\*assumption\*\*\*\*/

proc phreg data=cancer2 noprint;

class side(ref='1') gender(ref='0') race(ref='1') cat\_node cat\_ex;

model month\*status(0)= side gender race logage cat\_node cat\_ex logsize;

logage=log(agediag);

logsize=log(size);

output out = cancer1\_fitted logsurv = CoxSnell;

run;

/\*Cox-Snell residual\*/

data cancer1\_fitted;

set cancer1\_fitted;

CoxSnell = -CoxSnell;

label CoxSnell = 'Cox-Snell Residual';

run;

/\*Nelson-Aalen estimate\*/

proc phreg data = cancer1\_fitted noprint;

model CoxSnell\*status(0)= ;

output out = cancer1\_fitted2 logsurv = haz;

run;

data cancer1\_fitted2;

set cancer1\_fitted2;

haz = -haz;

label haz = 'Estimated Cumulative Hazard Rate';

run;

proc sort data = cancer1\_fitted2; by CoxSnell;

proc sgplot data = cancer1\_fitted2 noautolegend;

step x = CoxSnell y = haz;

lineparm x=0 y=0 slope=1;

run;/\*good enough\*/