Grade Effects on Survival in Breast Cancer Patients

Llasmin Lopez
Department of Statistics
University of California, Davis

The dataset gbsg.rda from the Rstudio data page resource was used to perform analysis.

This data set contains patient records from a trial conducted by the German Breast Cancer Study Group of 720 patients with node positive breast cancer. Only 686 patients have data for prognostic variables. The covariates are : patient id, age, menopausal status, tumor size, tumor grade, number of positive lymph nodes, progesterone receptors (fmol/l), estrogen receptors (fmol/l), hormonal therapy, recurrence free survival time; days to first of recurrence death or last follow-up, status. For this data, Time origin is 1984 as the study includes observations from 1984 to 1989 making the time scale on study time. The event of interest is recurrence or death. We can identify the mechanism of censoring and truncation as right censoring and left truncation. This is left truncated as patients who died before 1984 cannot be included in the study and right censored as the study concluded before all patients involved achieved the event of interest. We are interested in testing whether there is a difference between the survival curves stratified by grade of node and if grade is independent of another categorical variable, menopause. If lower grades have higher survival rates, we should place more effort into detecting cancer nodes early on.

Methods

Data was inspected with basic descriptive statistics to ensure data quality. We will provide descriptions of key concepts involved in this analysis. The Survival Function describes time-to-event phenomena by providing the probability of an individual surviving beyond time x. It is denoted as $S(x) = Pr(X > x)$.

The Hazard Function is defined by $h(x) = \lim_{\Delta x \to 0} \frac{P[x \leq X < x + \Delta x | X \geq x]}{\Delta x}$ or $h(x) = f(x)/S(x) = -d\log[S(x)]/dx$. if X is a continuous random variable. The Cumulative Hazard Function provides the total risk that has been accumulated at time x and is defined by $H(x) = \int_0^x h(u)du = -\log[S(x)]$. Note that the hazard function and cumulative hazard function involves the survival function.

The Kaplan- Meier estimator is used to estimate the survival function. This estimator is defined as follows for all values of t in the data where d_i is number of deaths and Y_i is the number of individuals at risk at time t_i

$$\hat{S}(t) = \prod_{t_i \leq t} [1 - \frac{d_i}{Y_i}], t_i \leq t$$

for $t_i \leq t$ and $\hat{S}(t) = 1$ for $t \leq t_1$.

The Cox Proportional Hazards Model, used to model the hazard function, is defined by

$$h(t|\mathbf{Z}) = h_0(t)c(\beta^t \mathbf{Z}) = h_0(t)\exp(\sum \beta_k \mathbf{Z}_k)$$

for $h_0(t)$ an arbitrary baseline hazard rate, $\beta = (\beta_1, ..., \beta_p)^t$ a parametric vector, and $c(\beta^t \mathbf{Z})$ a known function. This model assumes the observations are independent and identically distributed, censoring is independent, hazard ratio independent of time, and the hazard ratio for two Z's are proportional. The partial likelihood function is

$$L(\beta) = \prod_{i=1}^{D} \frac{\exp[\sum_k \beta_k Z_{(i)k}]}{\sum_{j \in R(t_i)} \exp[\sum_k \beta_k Z_{jk}]}$$

where $R(t_i)$ is the risk set at time t_i which includes all individuals who are still under study up to time t_i.

The Hazard Ratio compares two hazard rates and an estimate is provided by the Cox-Mantel estimate of HR for two groups

$$HR = \frac{H_A}{H_B} = \frac{O_A/E_A}{O_B/E_B}$$

where the observed number of events (deaths) in group i is O_i, the expected number of events (deaths) in group i is E_i, and H_i is the overall hazard rate for group i.

Hypothesis testing for the Cox PH model assumes a global null hypothesis $H_0 : \beta = \beta_0$ and the alternative global hypothesis $H_A : \beta \neq \beta_0$. We use the Wald test, likelihood ratio, and score test to test our hypothesis. The Wald test involves the partial likelihood function and

$$\chi^2_W = (\mathbf{b} - \beta_0)^t \mathbf{I}(\mathbf{b})(\mathbf{b} - \beta_0)$$

where $\mathbf{I}(\beta)$ is the information matrix. The Likelihood ratio test calculates the difference in -2log-likelihood

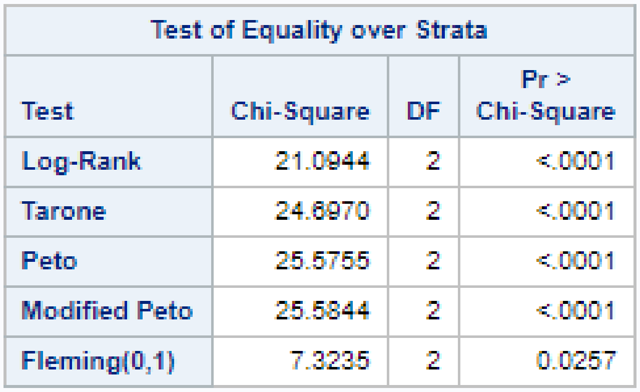
$$\chi^2_{LR} = 2[LL(\mathbf{b}) - LL(\beta_0)]$$

. Finally, the Score test derives the score equation and the information matrix.

To Test for Trend, an ordering of the hazard functions, we use the Logrank (Mantel-Haenszel) statistic where $X = \frac{(O_1-E_1)^2}{V_1} \sim \chi^2_1$ under the null hypothesis $H_0 : S_1(t) = S_2(t)$ for all t.

Results

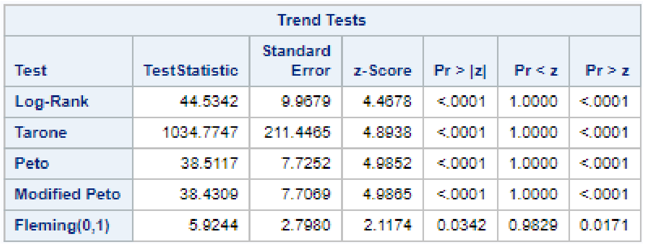
Upon first inspection using KM estimates, we see the survival curve for this data shows a decreasing survival rate as time increases. We found that about 56% of subjects were censored in this study, meaning they did not experience recurrence or death before 1989. Since we are particularly interested in the different grades, we run a K-sample test stratified by grade. The null hypothesis is $H_0 : h_1(t) = h_2(t) = h_3(t)$ for all $t \leq \tau$ and the alternative is H_A : at least one of the $h_j(t)$'s is different for some $t \leq \tau$



Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	21.0944	2	<.0001
Tarone	24.6970	2	<.0001
Peto	25.5755	2	<.0001
Modified Peto	25.5844	2	<.0001
Fleming(0,1)	7.3235	2	0.0257

Figure 1. K-sample test for strata of grade

It is clear to see that all of the tests in Figure 1 agree with the conclusion that the survival curves are not the same across the three grades. Since the grades are ordered by severity in real life, we will preform a test for trend. Here, the null is $H_0 : h_1(t) = h_2(t) = h_3(t)$ for all $t \leq \tau$ and the alternative is $H_A : h_1(t) \leq h_2(t) \leq h_3(t)$ for all $t \leq \tau$ with at least one strict inequality



Trend Tests						
Test	TestStatistic	Standard Error	z-Score	Pr > |z|	Pr < z	Pr > z
Log-Rank	44.5342	9.9679	4.4678	<.0001	1.0000	<.0001
Tarone	1034.7747	211.4465	4.8938	<.0001	1.0000	<.0001
Peto	38.5117	7.7252	4.9852	<.0001	1.0000	<.0001
Modified Peto	38.4309	7.7069	4.9865	<.0001	1.0000	<.0001
Fleming(0,1)	5.9244	2.7980	2.1174	0.0342	0.9829	0.0171

Figure 2. Trend test for strata of grade

From the table in Figure 2, we see strong support for rejecting the null hypothesis. It seems that at least one inequality holds in the alternative hypothesis.

We were also interested in the relationship between the covariates grade and menopause. Below, in Figure 3, we see the hazard curves stratified by grade and menopause. For menopause not present, the hazard ratios stratified by grade appear to cross, meaning an assumption has been violated.

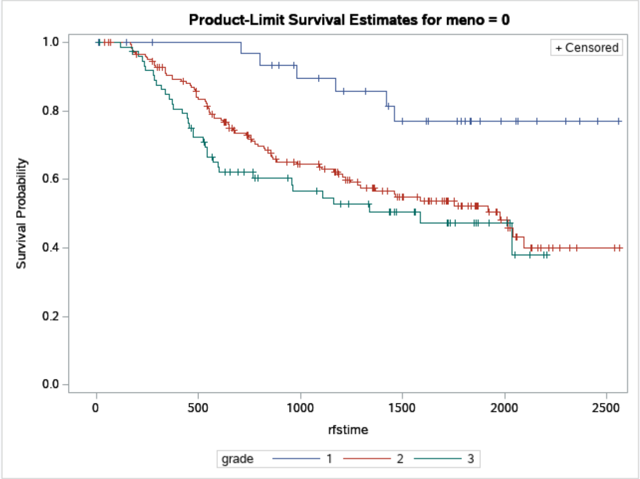


Figure 3. Trend test for strata of grade

The interaction between these variables should be further investigated before building a model of the hazard function.

Discussion

We have found sufficient evidence to conclude that grade level has an effect on patient survival time. We also found some interaction between grade and menopause. These two variables are related in some way by time: grade develops over time and menopause affects older women. This aligns with the general consensus of the field; the earlier a cancer is detected, the higher probability of survival.

Appendix:SAS

```
    LIBNAME proj1;
/** FOR CSV Files uploaded from Windows **/
FILENAME  CSV  "/home/u49996923/Project1/gbsg.csv"
TERMSTR=CRLF;

    /** Import the CSV file. **/

    PROC IMPORT DATAFILE=CSV
OUT=gbsg
DBMS=CSV
REPLACE;
getnames=yes;
RUN;

    /**Initial data checking and quality control**/
    * check variables;
proc contents data=gbsg;run;

    * Formatting variables;
proc format;
value fgrade
1=”1”
2=”2”
3=”3”
;
run;
proc format;
value fmeno
0=”0”
1=”1”
;
run;
proc format;
value fhormon
0=”0”
1=”1”
;
run;
proc format;
value fstatus
0=”0”
1=”1”
;
run;

    data gbsg;
set gbsg;
format grade fgrade.;
format meno fmeno.;
format hormon fhormon.;
format status fstatus.;
run;

    * check variables again;
proc contents data=gbsg;run;

    * select the first/last observation;
proc sort data=gbsg out=gbsg;
by id;
run;

    data first;
set gbsg;
by id;
if first.id=1;
run;

    * generate frequency table;
proc freq data=gbsg;
tables grade;
run;

    * check univariate data distribution;
proc univariate data=gbsg;
var rfstime age meno size grade nodes pgr er hormon status;
run;

    * check average;
proc means data=gbsg;
class grade;
var rfstime age meno size nodes pgr er hormon status;
run;

    /**Analyzing data**/

    * KM;
proc lifetest data=gbsg atrisk outs=KMest;
time rfstime*status(0);
run;

    proc lifetest data=gbsg atrisk outs=KMest;
time rfstime*status(0);
strata grade;
run;

    * K-sample test grade;
proc lifetest data=gbsg plots=survival(atrisk=0 to 2750 by
500);
time rfstime*status(0);
strata grade/test=(logrank TARONE PETO MODPETO
FLEMING(0,1) );
run;

    data covs;
format grade fgrade.;
input grade id;
datalines;
1 1
2 2
3 3
;
run;

    proc phreg data=gbsg plots(overlay)=(cumhaz);
class grade(desc);
model rfstime*status(0) = grade;
baseline covariates=covs out=base;
run;
proc phreg data =gbsg plots(overlay)=(survival);
class grade(desc);
model rfstime*status(0) = grade age meno size nodes pgr er
hormon;
strata grade;
run;

    * Nelson-Aalen;
proc lifetest data=gbsg nelson ;
time rfstime*status(0);
ods output productlimitestimates=ple;
run;
proc sgplot data = gbsg;
```

```
series x = rfstime y = CumHaz;

    * hazard function;
proc lifetest data=gbsg plots=hazard notable;
time rfstime*status(0);
run;

    *Stratefied test;
proc lifetest data=gbsg ;
time rfstime*status(0);
strata meno/group=grade test=(logrank TARONE PETO
MODPETO
FLEMING(0,1) );
run;


    * Tests for trend;
proc lifetest data=gbsg plots=survival(atrisk=0 to 2750 by
500);
time rfstime*status(0);
strata grade/trend test=(logrank TARONE PETO MODPETO
FLEMING(0,1) );
run;
```