**Workshop “Survival Analysis for Cancer Epidemiology”**

University of Nairobi - 21st & 22nd of October 2019

**Practical: Flexible Hazard-based Regression Models – Part 1 & 2**

Facilitators: Aurélien Belot and Bernard Rachet

Acknowledgement: We adapted the contents of a practical which was originally written by L.Remontet and M.Fauvernier (biostatistics department, Hospices Civils de Lyon, France) with the help of Z Uhry, E Dantony, H Charvat, A.Belot, N Bossard, L Roche and the CENSUR working survival group.

**Introduction**

For the 5 *practical works* (PW), we will use the “simulated” dataset of colon cancer patients diagnosed in England in 2000 and follow-up to the 31/12/2007 (file datacolon.Rdata in folder /data/). This dataset contains only male patients (n=6024).

The variables in this dataset are:

* *idenpat*: patients’ identification number
* *diagmdy*: date of diagnosis
* *agediag*: age at diagnosis (years)
* *stage*: stage at diagnosis (factor, St1=“early stage”, St2, St3, St4=“advanced stage”)
* *finmdy*: date of last know vital status
* *dead*: vital status (0=Alive, 1=Death)

Furthermore, the English population hazards are available in the file popex1.RData (folder /data/). The available variables are:

* *sex*: “Males” or “Females”
* *age*: age in years (from 0 to 100)
* *year*: from 1971 to 2009
* *rate*: mortality (all cause) hazard

**Practical work n°1**

Question 1.1:

load the data datacolon and describe it.

Question 1.2:

Using the function survfit of the package survival:

* Plot the survival by Kaplan-Meier (work in years for the follow-up)
* Plot the survival by Kaplan-Meier by stage.

Question 1.3:

For patient with stage 4 cancers (“St4”), calculate the mortality hazard over the all period (from 0 to 8 years), assuming a constant hazard.

Calculate the mortality hazard for

* the first 3 months (0-3),
* then for the 3 months that follow (3-6)
* then for the 6 months that follow (6-12)
* then by year of follow-up.

HINT: use the function lexis (in the folder /function/) with the following code to split the follow-up time for each patient:

lexis(entry=0, exit = fu, fail = dead, breaks =c(0, 0.25, 0.5, 1:8), data = datacolon, include=list(idenpat))

Plot these piecewise constant hazards *vs* time since diagnosis.

HINT: use the function segments

Note: the function lexis is a simpler (and older) version of function Lexis of package Epi.

**Practical work n°2**

Model: log(excess hazard) ~ piecewise constant

Question 2.1:

Load the population hazard from the file popex1 and plot the population hazard according to age for year 2000 and for Females and Males separately in log scale.

Question 2.2:

Merge the dataframe popex1 to the dataframe datacolon to retrieve the population hazard at the time of death or last known vital status (remember that datacolon contains only males patients) and save this new dataframe as datacolon.v2.Rdata

Question 2.3:

On the subset of patients diagnosed with stage 4 cancer from datacolon.v2, use the package mexhaz to fit a excess hazard model where the baseline is piecewise constant with intervals: "[0,0.25]" "(0.25,0.5]" "(0.5,1]" "(1,2]" "(2,3]" "(3,4]" "(4,5]". Call this model model23

Make a summary of model23, print the coefficients and the variance-covariance matrix.

Plot the excess hazard vs time since diagnosis from the estimated coefficients of model23

Save the values of the excess hazard in your working directory as hazard23.Rdata

Question 2.4:

From model23, calculate the net survival at 1 year and at 1.37 years “by hand” (i.e. using the coefficient of model23)

Check the results with predict.mexhaz (use predict(*mexhaz object*) )

Question 2.5:

Using the help below and the delta-method, calculate the confidence interval of these net survivals (at 1 year and 1.37 years), using a Wald-type CI for the cumulative hazard (i.e. assumption of normality for the cumulative hazard).

HELP:

1. var(exp(beta)) approximated with [exp(beta)]^2 \* var(beta)
2. model23$vcov is a diagonal matrix.

Check the results with predict.mexhaz

Question 2.6:

Using predict.mexhaz, compute the 95% confidence intervals for the Net survival at the same times assuming normality on log-log scale (i.e. : on the log-cumulative scale).

**Practical work n°3**

Model: log(excess hazard) ~ f(t)

Question 3.1:

On the subset of patients diagnosed with stage 4 cancer from datacolon.v2, use the package mexhaz to fit an excess hazard model (model31) with a cubic B-spline (one knot located 1 year) as baseline hazard and make hazard predictions for time t=seq(0,5,by=0.01) with predict.mexhaz

Hint: set the boundary with the command bound=c(0,8)

Question 3.2:

Check the fit done in 3.1 by 1) re-drawing the plot made in question 2.3 and 2) adding the predicted hazard from model31

Question 3.3:

Using the help below, calculate the predicted hazard from model31 "by hand".

HELP:

1. write the formulae of model31 in R terms (ie write the model as if you use bs() of library splines : as.formula("~bs(fu, k=1, B=c(0,8))" ))
2. use model.matrix and matrix product with coefficients of model31

Add this predicted hazard on the current graph (with line type lty=8).

Question 3.4:

Compute Net Survival at 1.5 years from model31 using Gauss-Legendre approximation (with 100 nodes) (try also with 20 nodes) and compare the results with predict.mexhaz;

HELP

1. use function gauss.quad of package statmod
2. GL <- gauss.quad(n=100,kind="legendre")
3. use the function

Rescale <- function(gl,a,b){gl$nodes <- gl$nodes\*(b-a)/2+(a+b)/2 ; gl$weights <- gl$weights\*(b-a)/2; return(gl) }

1. gg <- Rescale(GL,0,1.5)

**HELP Technical Note: Gauss-Legendre quadrature rule**

Approximates the integral of a function f over [−1, 1] by a weighted sum of function values at specified points (the nodes).

The higher the value of n, the better the approximation. In R, the function gauss.quad from the statmod library can be used to perform Gauss-Legendre (GL) quadrature rule. For example, the command gauss.quad(n=10,kind="legendre")$weights gives you the of the 10 points quadrature rule and gauss.quad(n=10,kind="legendre")$nodes gives you the .

An integral over [−1, 1] must be changed into an integral over [a, b] before applying the GL quadrature rule. This change of interval can be done in the following way:

Applying the GL quadrature rule then results in the following approximation:

Note: If you want to see a really easy method, make the calculs with “rectangle approximation” (however, not to do in real life)

Question 3.5:

Plot the hazard in black color using plot.predMexhaz (use plot(*mexhaz object*)); add confidence intervals in red

Plot the net survival in black color using plot.predMexhaz; add confidence intervals in red

Question 3.6:

Using the help below, check that the value of the likelihood of model31 does not depend on population mortality hazard of alive patients

HELP:

1. Change the values of the population mortality hazards for alive patients and check that it does not change the likelihood

**Practical work n°4**

**Model: log(excess hazard) ~ f(t)+ age\_class**

Question 4.1:

On the subset of patients diagnosed with stage 4 cancer and age [30;90] from datacolon.v2, fit an excess hazard model (model41) with a cubic B-spline (one knot located at 1 year) as baseline and a PH effect for the covariate age in 5 groups: "[30;45[", "[45;55[", "[55;65[","[65;75[", [75;90]" (reference="[65;75[")

Question 4.2:

Compute “by hand” (see question 3.3) the hazard at time 0:5 for the 5 age groups; check the results for age group [30;45[ and [65;75[ with predict.mexhaz

Question 4.3:

Compute the Hazard Ratio (HR) for each age groups (ref=[65;75[).

Given the model41, show that HR do not depend on the time since diagnosis (Proportional Hazard assumption)

Check that HR correspond to exp(model41$coef) for the corresponding age group

Question 4.4:

Plot the HR vs agediag

**Model: log(excess hazard) ~ f(t)+ g(age)**

Question 4.5:

Model the age effect with a cubic spline (with a knot at 70 years) (model 45) written in truncated powers basis (tips=centered and reduced agediag ie (agediag-70)/10)

Show that the log-likehood of model45 and the predicted are the same with bs basis (ie using bs(agediag,k=70))

Note that 70 is approximately the median(temp$agediag[temp$dead==1])

Question 4.6:

Add the HR (ref: age=70) from model45 to the current graph (done in 4.4)

**Model: log(excess hazard) ~ f(t)+ g(age) + age:h(t)**

Question 4.7:

Add a non-proportional effect of age to model45 (model47) and predict “by hand” the Excess\_hazard(t,age) for each combination of t=seq(0,5,by=0.1) , age=30:90

Notes: It is essential to check that the "by-hand" design matrix is correct by checking that the names of the coefficients (obtained with mexhaz) correspond with the name of the columns of this matrix (ie they are in the same order)

Question 4.8:

From model47, make figures:

1. hazard vs t for age=30,60,90 ==> fig1
2. hazard vs age for t=0.2, 0.5 1, 5 ==> fig2
3. a 3D plot hazard(t,a) vs (t,a) ==> fig3

Question 4.9:

From model47, make figures of the HR vs age at time 0.2, 0.5, 1, 5

Question 4.10:

From model47, calculate "by hand" the net survival at 2.1 years for individuals with the same age as idenpat n°8

Check the result with predict.mexhaz

Question 4.11:

Calculate the populational net survival by averaging the individual net survival

**Practical work n°5 (OPTIONAL)**

In this practical work you will have to fit 4 flexible excess hazard models by writing the likelihood yourself. Because this would be almost impossible to do it by yourself in the allocated time, we provided the R-code for convenience. So we advise you to look at the solutions and study them.

Question 5.0:

Load datacolon.v2.RData

**Model 1: log(excess hazard) = beta0**

Question 5.1:

Write the log-likelihood function for model1

Question 5.2:

Maximize the log-likelihood function for model1 using optim function (**with method=”Brent”**). Get estimation of beta0, the variance and the standard-error

Question 5.3:

Check your results with mexhaz package. Plot the hazard and the net survival

**Model 2: log(excess hazard) = beta0 + beta1\*time + beta2\*age**

Question 5.4:

Write the log-likelihood function. **Don’t forget to center the age variable to make sure your optimization succeeds**

Question 5.5:

Maximize the log-likelihood function using optim function, then get estimation of the parameters, the covariance matrix and the standard-errors

Question 5.6:

Check your results with mexhaz package. Plot the hazard and the net survival at ages 50, 60 and 70

**Model 3: log(excess hazard) = beta0 + Bspline(time, degree=3) + beta\*age**

Now things get really tricky since we no longer have a closed form of the cumulative excess hazard. You will have to use numerical integration techniques (Gauss-Legendre quadrature) using the gauss.quad function from the statmod library.

Question 5.7:

Define model 3 with a formula object. **The spline for time is without any knots**.

Question 5.8:

Get the Gauss-Legendre coefficients for [-1;1] with:

gauss.quad(n=10,kind="legendre")

Calculate the 10 design matrices that we will need to perform gauss-legendre quadrature :

Remember that we want to integrate exp(X%\*%beta). So we have to evaluate this expression at 10 different values of time to death. That’s why we need 10 design matrices X.

Use the technical Note on Gauss-Legendre to see how to deal with weights and nodes. The actual Gauss-Legendre will be performed inside the log-likelihood function at question 5.9.

Question 5.9:

Write the log-likelihood function. Use model.matrix to get the design matrix from the formula. **Don’t forget to center the age variable to make sure your optimization succeeds**

Question 5.10:

Maximize the log-likelihood function using optim function, then get estimation of the parameters, the covariance matrix and the standard-errors

Question 5.11:

Check your results with mexhaz package. Plot the excess hazard and the net survival at ages 50, 60 and 70

**Model 4: log(excess hazard) = beta0 + Bspline(time, degree=3) + beta\*age + Bspline(time, degree=3):age**

Question 5.12:

Define the model with a formula object. **The splines for time are without any knots**

Question 5.13:

Calculate the new 10 design matrices for Gauss-Legendre quadrature

Question 5.14:

Maximize the log-likelihood function using optim function, then get estimation of the parameters, the covariance matrix and the standard-errors. You can normally reuse the log-likelihood written for model 3.

Question 5.15:

Check your results with mexhaz package. Plot the excess hazard and the net survival at ages 50, 60 and 70