Unitat d'Estadística i Bioinformàtica (UEB)
Vall d'Hebron Institut de Recerca (VHIR)

Institut d'Investigació Sanitària acreditat per l'Instituto de Salud Carlos III (ISCIII)

Cutpoint determination in continous predictive variables in survival analysis

2014 Spanish Stata Users Group meeting

Barcelona, 23 Octubre 2014

Santiago Pérez Hoyos

santi.perezhoyos@vhir.org

Index

- Motivation
- How to select cutpoint
- Cavats
- Solutions
- Stata Flowchart
- Examples
- Further work

Motivation

- Increasing interest in use of binary categorization for continous variables involving clinical or epidemiological data (gene expressions, biomarkers, biochemical parameters,etc.)
- Main objective to build prognostic scores for a follow-up event
    - Easy to compute
    - Classify in high and low risk
    - Allows to calculate impact measures (hazard ratios)

Examples

- Survival 60 days after leaving ICU depending on severity scores and biochemical index
- Time to death or AIDS in HIV infected subjects depending on age, CD4 counts , RNA viral load, etc.
- Cancer survival depending on gene expression or biomarkers

How to select a cutpoint

- Based on graphics of the relationship of the continous variable with the outcome
- Median or percentiles
- Based on previous literature
- Based on best fit or the most significant relation with outcome (minimun p-value of all possible cutpoints)

Minimum p-value

- All values of prognostic variable except a proportion of extrem are cutpoints candidates
- Value that best separates outcomes according to maximum test statistics or minimum p-value is choosen
- Max Log-rank test is chosen in R Maxstat package
- Likelihood profile is chosen in our aproximation in Stata

Caveats

- Inflation or type I error rates
- Overestimate measures of effect
- Loss of information when categorizing
- Replicate the cutpoint in similar data

Solutions

- Bonferroni correction $P_{bonf} = p_{min} * n$ values
- considering data out P_5 and P_{95}
  - $P_{alt=}$ $-3.13\ p_{min}\ (1+1.65\ \ln(p_{min}))$
- Benjamini-Hochberg q values (qqvalue in stata)
- Cross-validation

Stata Template

- Define local parameters (regression type, titles, variables)
- Delete missing values, select data between percentiles 5 & 95
- Store null model
- Loop among unique values of quantitative variable and dichotomize variable
- Fit a regression model and calculate likelihood ratio test
- Save likelihood, p value in a temporary file
- Calculate Bonferroni and Benjamini-Hochberg corrections

Stata Template

- Select the minimum p-value (first obs after gsorting)
- Plot likelihood profile and hr profile
- Top ten table of p-values
- Fit regression model with selected cutpoint
- Use html functions for output
- Saving results & graphs in html file

Example

- Data from GEMES Spanish HIV seroconverters study
  - 2257 HIV seroconventers
  - Interested in time to death
  - CD4 and age as covariates
  - Find cutpoint for CD4 and age

Descriptiu cohort GEMES

	N individus	N Events	Taxa de incidencia*100 pers. Temps (I.C.95%)	Temps a risc
Total	2257	599	2.83 (2.61; 3.07)	21133.35

```
******* TROBAR PUNT DE TALL PER A VARIABLES CONTINUAS (CUTPOINT FINDER) ********
 local titol = "Temps a mort"   // Titol del gràfic0
 local tit_time = "Anys de seguiment" // Unitat de temps . Eix X
 local y_ord = "% de supervivents  " // Unitat de temps. Eix Y

**** Selecciona el tipus de regressio ******************************

local tiporeg="stcox"

if "`tiporeg'"=="logit" local RR="OR"  // Regressió logística
if "`tiporeg'"=="stcox" local RR="HR"  // Regressió de Cox
if "`tiporeg'"=="regress" local RR="B" // Regressió Lineal
if "`tiporeg'"=="poisson" local RR="RR" // Regressió Poisson

if "`tiporeg'"=="logit" local tr="exp" // Regressió logística
if "`tiporeg'"=="stcox" local tr="exp" // Regressió de Cox
if "`tiporeg'"=="regress" local tr="" // Regressió Lineal
if "`tiporeg'"=="poisson" local tr="exp" // Regressió Poisson0

*** Selecciona la variable resposta ******************************

local var_resp=""  // Variable resposta . En blanc per a Regressió de Cox
tempvar constant
gen `constant'=1

 *** Selecciona les variables del model per trobar punts de tall  **************
************************************************ CAL CANVIAR OOOJOOOOOO   ******

foreach vartalli in  cd4 edad {   // Variables quantitatives a dicotomizar
```

Top ten table

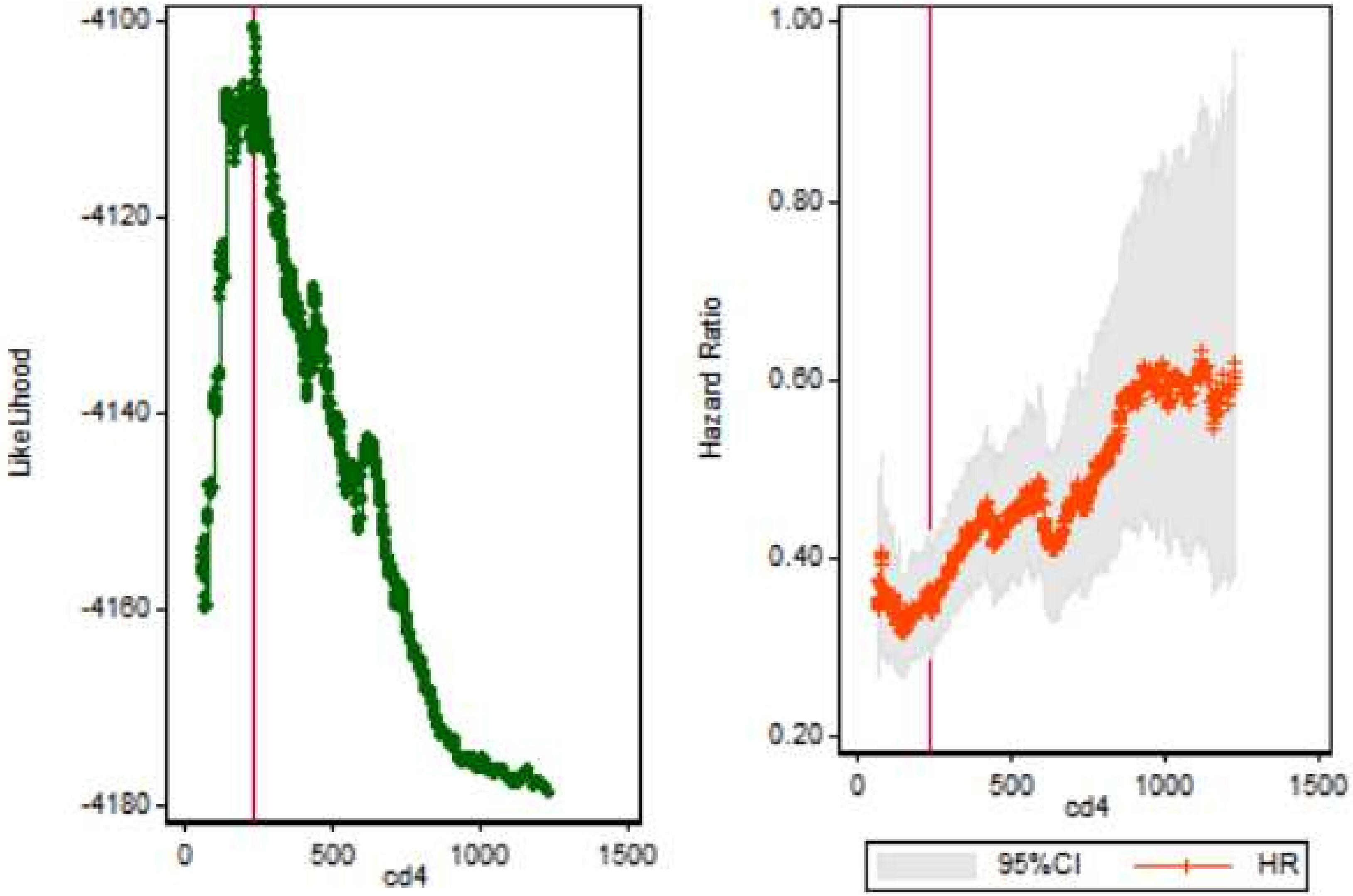
Anàlisi de supervivència Punts de tall de cd4

Top Ten Punts de tall de cd4

	P-value rank	Likelihood	Cutpoint	Hazard Ratio	Lower bound 95%CI	Upper bound 95%CI	p Value	p adjust Altman	Q value Benjamini-Hochberg	p adjust Bonferroni
1	1	-4100.598	240	0.34	0.29	0.40	5.89e-37	2.52e-34	3.338e-34	5.02e-34
2	2	-4100.882	241	0.34	0.29	0.40	7.84e-37	3.34e-34	3.338e-34	6.68e-34
3	3	-4101.718	242	0.34	0.29	0.40	1.82e-36	7.67e-34	3.439e-34	1.55e-33
4	4	-4101.774	243	0.34	0.29	0.40	1.92e-36	8.11e-34	3.439e-34	1.64e-33
5	5	-4101.822	244	0.34	0.29	0.40	2.02e-36	8.50e-34	3.439e-34	1.72e-33
6	6	-4102.542	245	0.34	0.29	0.40	4.16e-36	1.74e-33	5.909e-34	3.55e-33
7	7	-4104.054	246	0.35	0.30	0.41	1.91e-35	7.82e-33	2.188e-33	1.63e-32
8	8	-4104.128	246.5	0.35	0.30	0.41	2.05e-35	8.41e-33	2.188e-33	1.75e-32
9	9	-4105.076	247	0.35	0.30	0.41	5.33e-35	2.16e-32	5.048e-33	4.54e-32
10	10	-4105.958	248	0.35	0.30	0.41	1.30e-34	5.18e-32	1.104e-32	1.10e-31

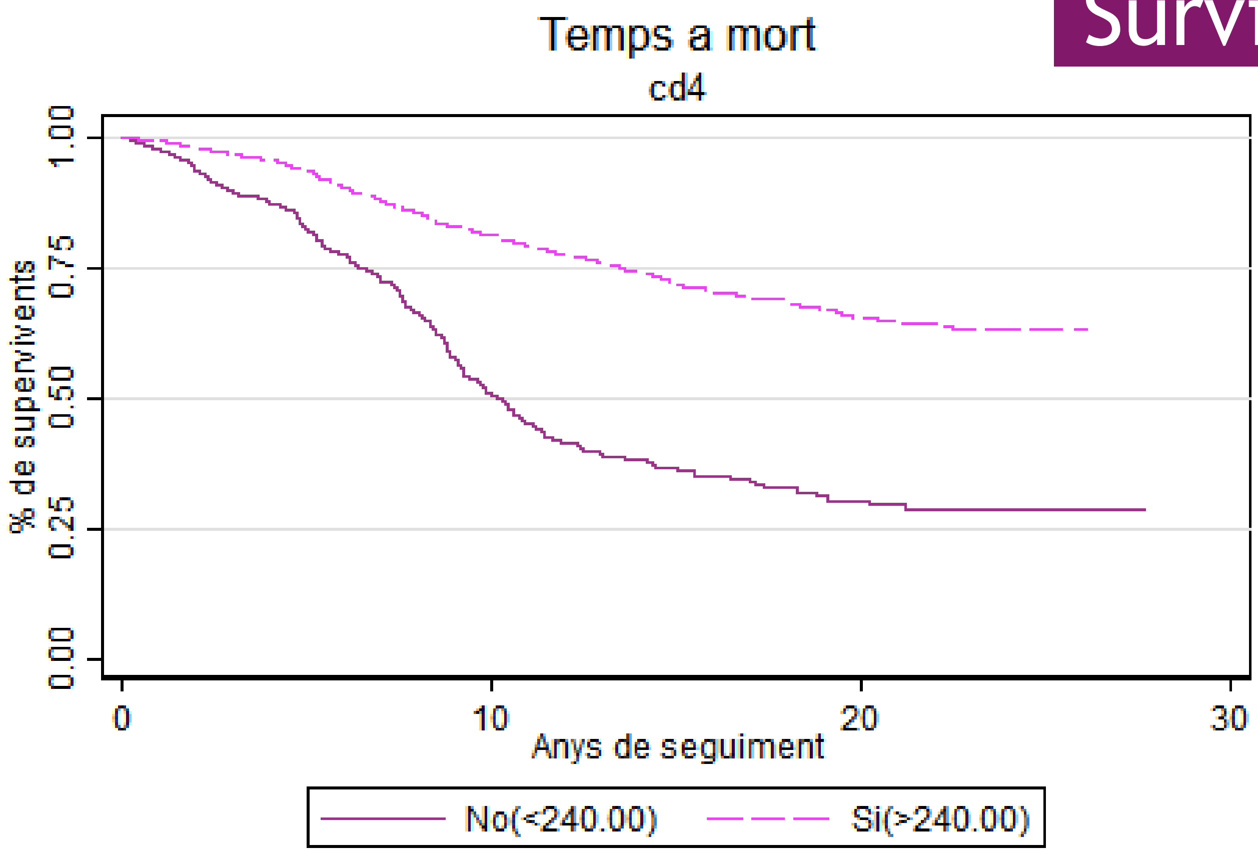
Profile plot

Punts de tall per cd4 en 240.00 (cd4 >240.00)



valor optimo 240.00, q value 0.0000

Survival after cutpoint



P valor = 0.0000 Bonferroni= 0.0000 Q valor =0.0000

cd4 >240.00	N individus	N Events	Taxa de incidencia*100 pers. Temps (I.C.95%)	Temps a risc	Temps Q_{25}	Temps Mediana (Temps)	Temps Q_{75}	P valor
No(<240.00)	491	274	5.84 (5.17; 6.57)	4693.86	6.45	10.16	.	
Si(>240.00)	1762	322	1.96 (1.76; 2.19)	16393.95	13.47	.	.	0.0000
Total	2253	596	2.83 (2.60; 3.06)	21087.81	9.32	.	.	

Cox after cutpoint

Regressio multivariant stcox

Number of obs = 2253

VARIABLE		HR	(95%CI)	p-value
cd4 >240.00	No(<240.00)	1		0.0000
	Si(>240.00)	0.34	(0.29; 0.40)	

LL model= -4100.60 ; AIC model= 8203.20 ; BIC model= 8208.92

p valor ajustado 0.0000

Regressio multivariant stcox. p valors ajustats

Number of obs = 2253

VARIABLE		HR	(95%CI)	p-value
cd4 >240.00	No(<240.00)	1		0.0000
	Si(>240.00)	0.34	(0.29; 0.40)	

valors p ajustats Bonferroni

Regressio multivariant stcox. p valors ajustats

Number of obs = 2253

VARIABLE		HR	(95%CI)	p-value
cd4 >240.00	No(<240.00)	1		0.0000
	Si(>240.00)	0.34	(0.29; 0.40)	

valors p ajustats Altman

Survival after cutpoint

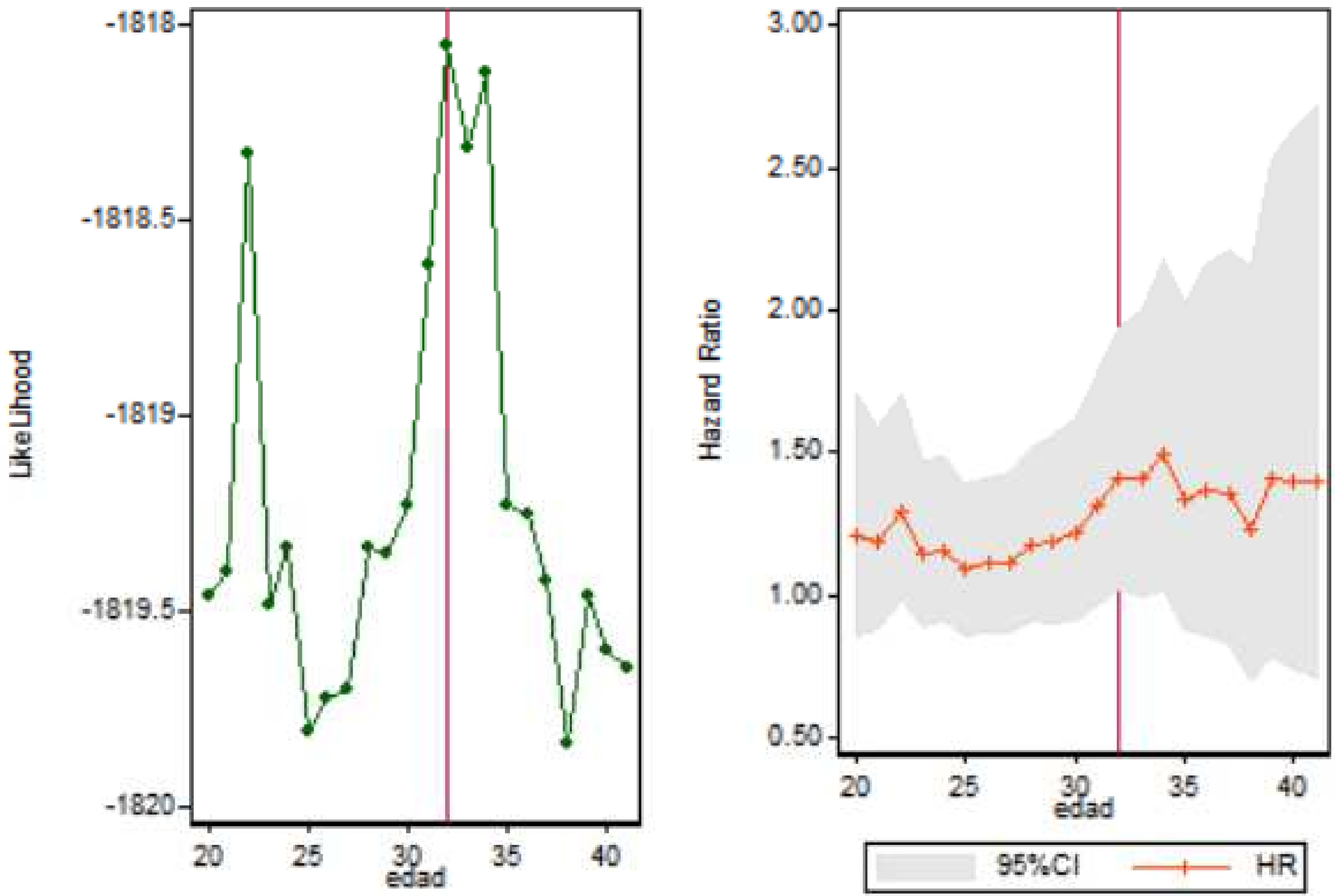
Anàlisi de supervivència Punts de tall de edad

Top Ten Punts de tall de edad

	P-value rank	Likelihood	Cutpoint	Hazard Ratio	Lower bound 95%CI	Upper bound 95%CI	p Value	p adjust Altman	Q value Benjamini-Hochberg	p adjust Bonferroni
1	1	-1818.052	32	1.41	1.02	1.94	.0440258	.572275	.33807288	.9685677
2	2	-1818.123	34	1.49	1.02	2.18	.0478806	.6016266	.33807288	1
3	3	-1818.315	33	1.41	1.00	2.00	.0602932	.685816	.33807288	1
4	4	-1818.331	22	1.29	0.98	1.70	.0614678	.6930518	.33807288	1
5	5	-1818.612	31	1.31	0.97	1.78	.0866564	.8233535	.37960526	1
6	6	-1819.226	35	1.33	0.88	2.02	.1912814	1.03524	.37960526	1
7	7	-1819.232	30	1.22	0.91	1.62	.1929407	1.035614	.37960526	1
8	8	-1819.252	36	1.37	0.87	2.16	.1983415	1.036324	.37960526	1
9	9	-1819.336	28	1.18	0.91	1.52	.2226326	1.030406	.37960526	1
10	10	-1819.338	24	1.16	0.91	1.48	.2232034	1.030096	.37960526	1

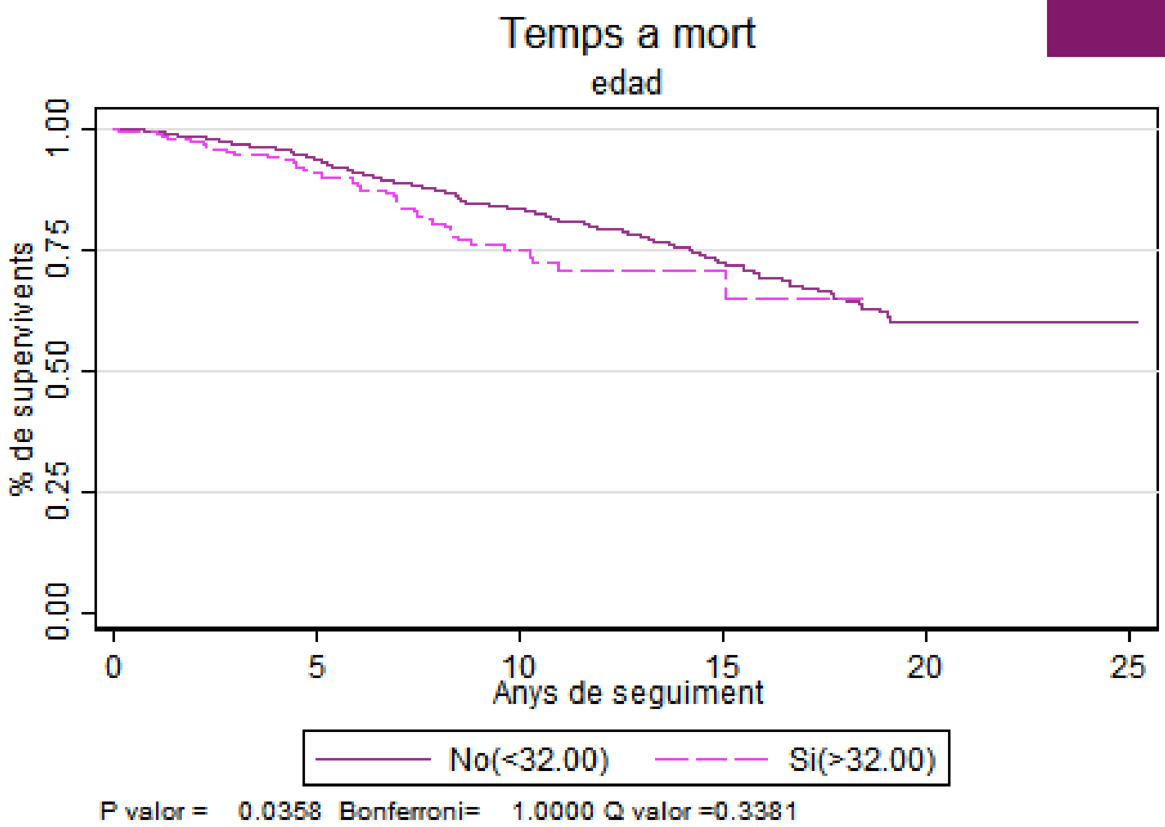
Profile plot

Punts de tall per edad en 32.00 (edad >32.00)



valor optimo 32.00, q value 0.3381

Profile plot



edad >32.00	N individus	N Events	Taxa de incidencia*100 pers. Temps (I.C.95%)	Temps a risc	Temps Q_{25}	Temps Mediana (Temps)	Temps Q_{75}	P valor
No(<32.00)	1332	230	1.96 (1.72; 2.23)	11717.50	14.20	.	.	0.0358
Si(>32.00)	363	46	2.39 (1.75; 3.19)	1924.95	9.58	.	.	
Total	1695	276	2.02 (1.79; 2.28)	13642.45	13.64	.	.	

Cox after cutpoint

Regressio multivariant stcox

Number of obs = 1695

VARIABLE		HR	(95%CI)	p-value
edad >32.00	No(<32.00)	1		0.0367
	Si(>32.00)	1.41	(1.02; 1.94)	

LL model= -1818.05 ; AIC model= 3638.10 ; BIC model= 3643.54

p valor ajustado 0.3381

Regressio multivariant stcox. p valors ajustats

Number of obs = 1695

VARIABLE		HR	(95%CI)	p-value
edad >32.00	No(<32.00)	1		0.8077
	Si(>32.00)	1.41	(1.02; 1.94)	

valors p ajustats Bonferroni

Regressio multivariant stcox. p valors ajustats

Number of obs = 1695

VARIABLE		HR	(95%CI)	p-value
edad >32.00	No(<32.00)	1		0.5117
	Si(>32.00)	1.41	(1.02; 1.94)	

valors p ajustats Altman

Further work

- Build an ado function in Stata
- Extent to more than one variable
- Extent to more than one cutpoint

- Mazudar M, Glassman JR. Tutorial in Biostatistics: Categorizing a prognostic variable review of methods. Code for easy implementation and applications to decision making about Cancer treatments. Sata in Med.2000; 19: 113-132
- William BA, Mandrekar JN, Mandrekar SJ, Cha SS, Furth AF. Finding optimal cutpoints for continous covariates with binary and time-to-event outcomes. Techical Report Series #79. Department of Health Scientes Research Mayo Clinic. 2006.
- Hothorn T, Lausen B. Maximally selected rank statistics with several p-value approximations. R Package 'maxstat' July 2, 2014

Thanks

Gràcies

Gracias