Gustavo Brusse

EDSD 2018/2019

Assignment – Event History 2

Exercise 1

Problem 1.1

First, we should split time axis into M + 1 pre-defined intervals: (0, 0.25, 0.5, 1, 2, 3). Using the command “survSplit”, it is possible to split the time axis and replicate the individuals by their respective intervals participation. Here are the first 6 observations of the dataset after procedure:

```
ID sex marstat   agegr health start  stay event interval
 1  1   f       0 (75,90]     1  0.00 0.101     1        1
 2  2   f       0 (75,90]     2  0.00 0.167     1        1
 3  3   f       0 [65,75]     2  0.00 0.250     0        1
 4  3   f       0 [65,75]     2  0.25 0.500     0        2
 5  3   f       0 [65,75]     2  0.50 1.000     0        3
 6  3   f       0 [65,75]     2  1.00 2.000     0        4
```

Problem 1.2

Piece-wise constant hazard model can be written by:

$$ln\mu_{ij} = lny_{ij} + \alpha_j$$

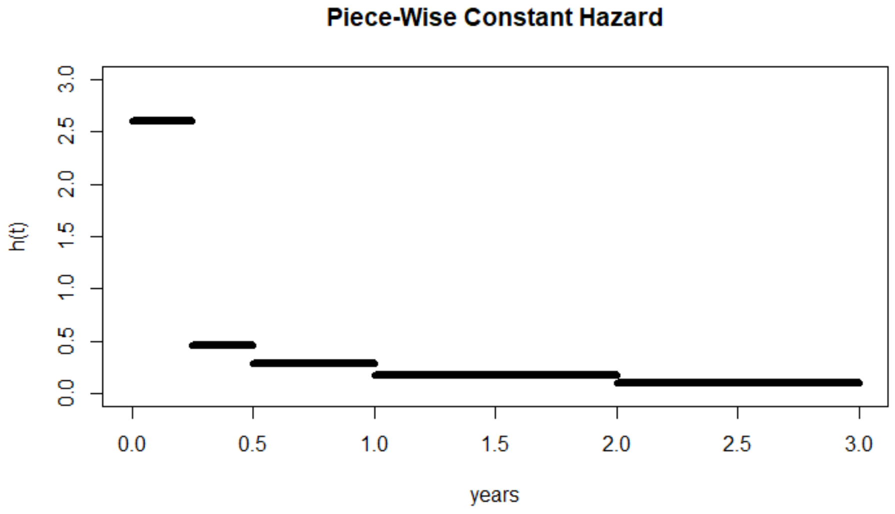
For each interval j we have one α intercept and the time spent by each individual in each interval as an offset (via the log-link) lny_{ij}. The outcome with the estimate intercept coefficients are:

Intercepts	Estimate Values
α_1	0.95
α_2	-0.78
α_3	-1.26
α_4	-1.79
α_5	-2.36

All the coefficients are significant at 0.001. The statistics significance of all intercepts means that we reject the null hypothesis that the hazard is constant over time. So, for each interval we have significant different hazard as baseline.

Problem 1.3

The Plot of the resulting piece-wise constant hazard is show below.



The graph shows that the longer you stay in nursing home, less likelihood one experience the event "returning to home".

Problem 1.4

Comparing all the models by an Analysis of Deviance qui-squared test, the model that is most suitable for describing the is the one with Sex, Marital Status and Health covariates:

$$mSMH : hi(t) = h_0(t) \cdot exp(\beta_1 \cdot sex + \beta_2 \cdot marstat + \beta_{3,4} \cdot health)$$

The models mS and mSMHA were not statistically significant. Comparing mSM and mSMH model, the addition of "health" variable produces a reduction in Deviance. Deviance is a measure that describes the badness of fit. So, the higher numbers indicate the worse is the fit. In conclusion, mSMH is considered the best model.

Exercise 2

As the information on survival times of 686 patients with primary node positive breast cancer is left censured, we calculate the log-likelihood as:

$$l(a,b,\beta_1,\beta_2) \ = \sum_{i}^{n} \delta_i(\ln\left(h_0(y_i)\right) + \boldsymbol{X\beta}) - (H_0(y_i) * e^{\boldsymbol{X\beta}})$$

Where δ_i is the dummy variable that indicate if the patience exit was due to death, $h_0(y_i)$ is the baseline hazard given by a Weibull distribution with parameters a and b, and $\boldsymbol{X\beta}$ is the regression part that consider the covariates and the respective parameters.

By the maximum likelihood estimation method, the estimated proportional hazard model when the only available covariate is the prognostic group is given by:

$$h_i(t) \ = \frac{1.38}{3.40}\left(\frac{t}{3.40}\right)^{0.38} * e^{\boldsymbol{X\beta}}$$

Where

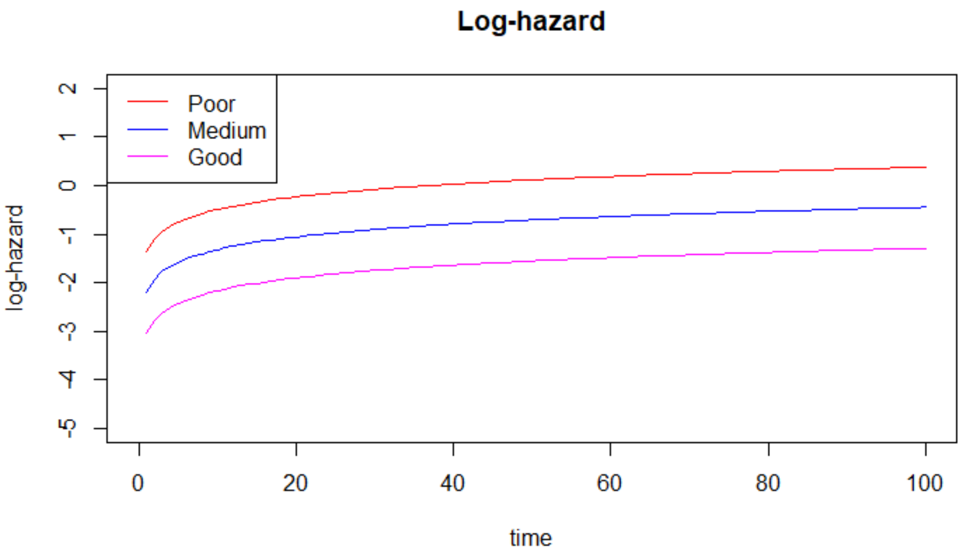
$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \ and \ \boldsymbol{\beta} = \begin{bmatrix} -1.27 \\ -0.83 \end{bmatrix}$$

The respective estimated parameters and their 95% confidence interval are show below:

Parameter	**Estimated Parameter**	**IC Lower boundary**	**IC upper boundary**
$\boldsymbol{\beta_1}$	-1.67	-1.99	-1.35
$\boldsymbol{\beta_2}$	-0.83	-1.08	-0.57
a	1.38	1.25	1.51
b	3.40	3.00	3.80

As the estimated parameters $\widehat{\boldsymbol{\beta_1}}$ and $\widehat{\boldsymbol{\beta_2}}$ are statistically different (95% confidence interval doesn't contain 0), we can say that the prognostic groups are different.

The log-hazard for each of the three prognostic groups are shown below. It can be seen that if a patient prognostic was "good", they had less likelihood to leave the study by death than those patients whose prognostic was "medium" or "poor". Those patients whose prognostic was "poor" had the highest likelihood to leave the study by death. Also, we can see that the log-hazard increase during the time for all categories.



Exercise 3

1) Without knowing anything about the baseline hazard, the Cox PH model can be written just as:

$$h_i(t) = e^{X\beta}$$

Including only gender, residence and gender/residence interaction we have:

$$h_i(t) = exp\left(X_{gender}\beta_{gender} + X_{urbrur}\beta_{urbrur} + X_{gender} * X_{urbrur}\beta_{gender/urbrur}\right)$$

The respective estimated parameters, their 95% confidence interval and significance are show below:

Parameter	Exp.(coefficients)	IC Lower boundary	IC upper boundary	P-value
$\boldsymbol{\beta_{gender}}$	0.89	0.80	0.99	0.03
$\boldsymbol{\beta}_{urbrur}$	1.10	1.00	122	0.03
$\boldsymbol{\beta_{gender/urbrur}}$	0.93	0.82	1.05	0.25

a) Women had 11% less chance to leave the Chinese Longitudinal Healthy Longevity Survey by death compared to man.

b) The interaction is not significant at a $\alpha = 0.05$. We should not consider the interaction of gender and residence to evaluate the surviving of this population. So, it doesn't matter if the women live in rural area or urban area, they will have less chances of dying than man anyway.

2) Using the stepwise method, the model that best fit the data according to AIC criterion is:

$$h_i(t) = exp(X_{gender}\beta_{gender} + X_{urbrur}\beta_{urbrur} + X_{act}\beta_{act} + X_{adl}\beta_{adl} + X_{urbrur} * X_{act}\beta_{act/urbrur} + X_{gender}\beta_{adl} * X_{act}\beta_{adl/gender})$$

Covariates: gender, urbrur, act, adl, urbrur*act, gender*adl

The respective estimated parameters, their 95% confidence interval and significance are show below:

Parameter	Exp.(coefficients)	IC Lower boundary	IC upper boundary	P-value
$\boldsymbol{\beta}_{\boldsymbol{gender}}$	0.81	0.75	0.88	0.00
$\boldsymbol{\beta}_{urbrur}$	1.01	0.93	1.08	0.79
$\boldsymbol{\beta}_{act}$	0.62	0.56	0.69	0.00
$\boldsymbol{\beta}_{adl}$	1.32	1.15	1.52	0.00
$\boldsymbol{\beta}_{adl}$	2.07	1.85	2.32	0.00
$\boldsymbol{\beta}_{\boldsymbol{urbur/act1}}$	1.17	1.03	1.33	0.01
$\boldsymbol{\beta}_{\boldsymbol{gender/adl1}}$	0.93	0.78	1.11	0.43
$\boldsymbol{\beta}_{\boldsymbol{gender/adl2}}$	0.83	0.73	0.96	0.01

The only significant interactions coefficients terms at $\alpha = 0.05$ were $\boldsymbol{\beta}_{\boldsymbol{urbur/act1}}$ and $\boldsymbol{\beta}_{\boldsymbol{gender/adl2}}$. It means that the condition effect of living in rural area and being active is more than being urban and active. Also, the condition effect of being women and having with two or more limitation is less than being a man with the same conditions

3) Using the estimate model, we can predict:

Estimated survival function of a sedentary male to a female sedentary, both living in a rural area, with no limitations in activities of daily living:

$Hazard\ Ratio$:

$$= \frac{exp\left(1_{gender}\beta_{gender} + 1_{urbrur}\beta_{urbrur} + 0_{act}\beta_{act} + 0_{adl}\beta_{adl} + 1_{urbrur} * 0_{act}\beta_{act/urbrur} + 0_{gender}\beta_{adl} * 0_{act}\beta_{adl/gender}\right)}{exp\left(0_{gender0}\beta_{gender} + 1_{urbrur}\beta_{urbrur} + 0_{act}\beta_{act} + 0_{adl}\beta_{adl} + 1_{urbrur} * 0_{act}\beta_{act/urbrur} + 1_{gender}\beta_{adl} * 0_{act}\beta_{adl/gender}\right)}$$

$$= \frac{exp\left(1_{gender}\beta_{gender} + 1_{urbrur}\beta_{urbrur}\right)}{exp\left(0_{gender}\beta_{gender} + 1_{urbrur}\beta_{urbrur}\right)} = \frac{\exp(-0.20) * \exp(0.01)}{\exp(0.01)} = 0.81$$

A sedentary female, living in a rural area, with no limitations in activities of daily living has 19% less chance to die compare to man with the same conditions.