Homework 01

ID: 1831

1. **Methods**: time to death unity is converted to years by dividing the original variable by 365.25 and then minimal time is found for those known to be alive on last follow up.

**Results**: The minimal time of follow up on those known to be alive on the last follow up visit is 4.05 years which is a little bit more than 4 years. And all records on the dataset have information on time to death. So it is appropriate to dichotomize the time according to death within 4 years of study or death after 4 years.

2. **Methods**: In order the describe baseline characteristics per levels of blood C reactive protein (CRP), this variable has been categorized in 3 levels according to recommended levels for cardiovascular disease (below 1, 1 – 3 and above 3 mg/L). Then per each level of the categorized CRP descriptives (mean, standard deviation, minimum and maximum) are used for continuous variables (age, serum cholesterol and body mass index) and proportions are used for the binary variables (gender, current smoking and prior atherosclerotic disease).

The dataset contained 5000 records but 89 (1.8%) are omitted from the analysis due to missing value (67 missed CRP value and 22 had a missing value in any of the other variables of interest).

**Results**: Only 4911 records are used: 426 had CRP below 1 mg/l, 2615 between 1 and 3 mg/l and 1870 above 3 mg/l. Table 1 presents descriptive statistics within these groups. Subjects having in the lowest interval CRP were more likely to be male than in other intervals. No consistent trend was found on the mean age at recruitment. On the other hand the mean BMI, prevalence of smokers, mean levels of serum cholesterol and prevalence of atherosclerotic disease increase with increase of levels of CRP.

 Table 1 – Participants characteristics at enrollment according to levels of CRP

|  |  |
| --- | --- |
|  | **Blood C reactive protein levels (mg/l)** |
| **Characteristic** | **< 1** | **1 - 3** | **≥ 3** | **Total** |
| N = 426 | N = 2615 | N = 1870 | N = 4911 |
| Male (%) | 45.5 | 44.6 | 37.8 | 42.0 |
| Age (years)\* | 73.4 (5.79; 65 - 94) | 72.8 (5.58; 65 - 100) | 72.6 (5.49; 65 - 93) | 72.8 (5.57; 65 - 100) |
| BMI (Kg/m2)\* | 23.8 (3.64; 15.6 - 38.6) | 26.1 (4.17; 14.7 - 53.2) | 28.1 (5.18; 15.3 - 58.8) | 26.7 (4.72; 14.7 - 58.8) |
| Current smoker (%) | 9.6 | 10.1 | 15.7 | 12.2 |
| Serum cholesterol (mg/dl)\* | 206.1 (40.46; 109 - 407) | 212.5 (38.65; 73 - 363) | 211.8 (39.70; 96 - 430) | 211.7 (39.24; 73 - 430) |
| Prior atherosclerotic disease (%) | 18.3 | 20.7 | 27.2 | 22.9 |
|   |   |   |   |   |
| \* Descriptive statistics: mean (standard deviation; minimum - maximum) |

3. **Methods**: Mean values of serum CRP are compared between subjects who died within first 4 years of study follow up and those who survived at least 4 years. T-test with unequal variances is used to test the test mean difference to be different from zero (bilateral test with 5% level). Also 95% confidence intervals (95% CI) are presented.

The dataset used for this analysis contains 4911 records as explained on question 2 methods section.

**Results**: The mean serum level of CRP was 5.4 mg/l (95% CI: 4.7 – 6.1) among the 482 subjects dying within 4 years of enrollment compared to 3.4 mg/l (95% CI: 3.2 – 3.6) among 4429 subjects surviving at least these first 4 years after enrollment. There is a tendency of 2 mg/l higher CRP levels among the subjects dying within 4 years of enrollment and it wouldn’t be unusual to find this difference between 1.2 and 2.7 mg/l (p-value < 0.0001).

4. **Methods**: Geometric mean values of serum CRP are compared between subjects who died within first 4 years of study follow up and those who survived at least 4 years. T-test with unequal variances is used to test the test mean log difference to be different from zero (bilateral test with 5% level). Also 95% confidence intervals (95% CI) are presented.

The CRP values contain 0 so before computing the logarithms 1 is added to all CRP values. Then the mean logs are retransformed by exponentiation and subtraction of 1.

The dataset used for this analysis contains 4911 records as explained on question 2 methods section.

**Results**: The geometric mean serum level of CRP was 3.27 mg/l (95% CI: 2.97 – 3.59) among the 482 subjects dying within 4 years of enrollment compared to 2.18 mg/l (95% CI: 2.11 – 2.24) among 4429 subjects surviving at least these first 4 years after enrollment. There is a tendency of 34.4% higher CRP levels among the subjects dying within 4 years of enrollment compared to those who survived and it wouldn’t be unusual to find this relative increase between 25.5 and 43.9% (p-value < 0.0001).

5. **Methods**: the proportion of subjects dying within 4 years of study enrollment is compared between subjects who had serum levels of CRP greater or equal to 3mg/l (higher levels) and subjects whose serum levels of CRP were below 3mg/l (lower levels). Then a difference on these proportions (risk difference) is computed and tested using Pearson’s Chi squared test for independence. 95% confidence intervals of the risk difference are also presented.

The dataset used for this analysis contains 4911 records as explained on question 2 methods section.

**Results**: Among the 3041 subjects with lower levels of serum CRP 7.3% died within first 4 years of enrollment whereas among the 1870 subjects with higher levels of serum CRP 14.0% died over the same time of follow up. This gives an excess risk among the subjects with higher serum levels of CRP of 6.7% which wouldn’t be unusual to lay between 4.9% and 8.5% (with p-value < 0.0001).

6. **Methods**: the odds of subjects dying within 4 years of study enrollment is compared between subjects who had serum levels of CRP greater or equal to 3mg/l (higher levels) and subjects whose serum levels of CRP were below 3mg/l (lower levels). Then odds-ratio is computed and tested to be different from 1 using Pearson’s Chi squared test for independence. 95% confidence intervals of the odds-ration are also presented.

The dataset used for this analysis contains 4911 records as explained on question 2 methods section.

**Results**: Among the 3041 subjects with lower levels of serum CRP the odds of dying within first 4 years of enrollment is 0.0784 whereas among the 1870 subjects with higher levels of serum CRP the odds of dying within first 4 years of enrollment was 0.1622. So the odds of dying within 4 years of enrollment is 2.07 times higher among the subjects with higher serum levels of CRP compared to those with lower levels of CRP and it wouldn’t be unusual to lay between 1.71 and 2.50 (with p-value < 0.0001).

7. **Methods**: the survival distribution was estimated using Kaplan-Meier estimates in each group of CRP levels defined as below 3 mg/l (lower) and greater or equal than 3 mg/l (higher). Difference in survival distributions between these groups was tested using logrank test and bilateral p-value is reported. The hazard ratio and 95% confidence interval was computed using Cox proportional hazards regression with a single predictor (the CRP level).

The dataset used for this analysis contains 4911 records as explained on question 2 methods section.

**Results**: Graph 1 depicts Kaplan-Meier estimates of survival probability for the 3041 subjects whose serum levels of CRP is lower compared to 1870 subjects at the higher CRP serum levels. The graph shows the tendency of lower survival among those on higher category of CRP serum levels at every point in time. The instantaneous risk of death is estimated to be 60.1% higher on the group with serum levels of CRP above or equal to 3mg/dl and this estimate would not be unusual to be found between 42.3 – 80.2%. A logrank test two-sided p value of 0.0001 suggests that the null hypothesis of no survival difference to be rejected.

Graph 1 – Kaplan-Meier survival estimates

8. Any of the analysis would be valid as long they are specified *a priori*. It is important to notice that most times is not a good idea to dichotomize a continuous variable as we did with CRP due to loss of precision. Given the suggestion of logarithm of CRP as predictor a multiplicative model would preferred even though the proportions are much easier to understand than the hazard-ratio or odds-ratio. The hazard-ratio considers the exact (closer) time to death and the technique used manages the censoring. So I would prefer the hazard-ratio and Cox regression.

**APPENDIX**

**Stata Code**

clear

/\*

infile id site age male bkrace smoker estrogen prevdis diab2 bmi ///

 systBP aai cholest crp fib ttodth death cvddth ///

 using http://www.emersonstatistics.com/datasets/inflamm.txt

drop in 1

save inflamm.dta, replace

\*/

// import delimited using http://www.emersonstatistics.com/datasets/inflamm.txt

use inflamm.dta

// Question 1

g ttodthyr = ttodth/365.25

// Just use the minimal time on the alive

tabstat ttodthyr, stat(min n) by(death)

gen deadin4 = (death==1 & ttodthyr < 4) if ttodthyr < .

// Question 2

sum crp // There are 67 records withouth crp value: 11 died within 4 years

egen crpcat = cut(crp), at(0 1 3 110) icodes label

keep if crpcat < . // Drop if no CRP info

egen miss = rowmiss(age male bmi smoker cholest prevdis) // 22 miss any of these variables

keep if miss == 0

count

di (5000 - r(N))/5000

ctabstat age male bmi smoker cholest prevdis, stat(mean sd min max N) by(crpcat) ///

 col(var) long sep

// Question 3

ttest crp, by(deadin4) unequal

// Question 4

gen logcrp = log(crp + 1)

ttest logcrp, by(deadin4)

di "mean 1 =" %4.2f exp(r(mu\_1)) - 1

di "CI 95% 1: " %4.2f exp(r(mu\_1) - invnormal(0.975)\*r(sd\_1)/sqrt(r(N\_1) - 1)) - 1 ///

 "-" %4.2f exp(r(mu\_1) + invnormal(0.975)\*r(sd\_1)/sqrt(r(N\_1) - 1)) - 1

di "----"

di "mean 2 =" %4.2f exp(r(mu\_2)) - 1

di "CI 95% 2: " %4.2f exp(r(mu\_2) - invnormal(0.975)\*r(sd\_2)/sqrt(r(N\_2) - 1)) - 1 ///

 "-" %4.2f exp(r(mu\_2) + invnormal(0.975)\*r(sd\_2)/sqrt(r(N\_2) - 1)) - 1

di "----"

di "Ratio increase (%) =" %4.1f 100\*(exp(r(mu\_2) - r(mu\_1)) - 1) "%"

di "CI 95%: " %4.1f 100\*(exp(r(mu\_2) - r(mu\_1) - invnormal(0.975)\*r(se)) - 1) ///

 "-" %4.1f 100\*(exp(r(mu\_2) - r(mu\_1) + invnormal(0.975)\*r(se)) - 1) "%"

// Question 5

egen crpcat2 = cut(crp), at(0 3 110) icodes label

tab deadin4 crpcat2, col

cs deadin4 crpcat2

// Question 6

cc deadin4 crpcat2

tabodds deadin4 crpcat2, or

di 221/2820

di 261/1609

// Question 7

stset ttodth, failure(death) scale(365.25)

stcox crpcat2

sts test crpcat2

sts graph, by(crpcat2) ///

 risktable(, order(1 "Lower CRP level" 2 "Higher CRP level") righttitles) ///

 xtitle(Time after enrolment (years)) ///

 ytitle(Survival probability) ylabel(0(0.1)1, angle(horizontal) format(%2.1f)) ///

 xlabel(0 (1) 8) ///

 legend(order(1 "Lower CRP level" 2 "Higher CRP level"))

graph export survival.tif, replace