**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2015

**Homework #1**

January 5, 2015

**Written problems:** To be submitted as a MS-Word compatible file to the class Catalyst dropbox by 9:30 am on Monday, January 12, 2015. See the instructions for peer grading of the homework that are posted on the web pages.

*On this (as all homeworks) Stata / R code and unedited Stata / R output is* ***TOTALLY*** *unacceptable. Instead, prepare a table of statistics gleaned from the Stata output. The table should be appropriate for inclusion in a scientific report, with all statistics rounded to a reasonable number of significant digits. (I am interested in how statistics are used to answer the scientific question.)*

***In all problems requesting “statistical analyses” (either descriptive or inferential), you should present both***

* ***Methods: A brief sentence or paragraph describing the statistical methods you used. This should be using wording suitable for a scientific journal, though it might be a little more detailed. A reader should be able to reproduce your analysis. DO NOT PROVIDE Stata OR R CODE.***
* ***Inference: A paragraph providing full statistical inference in answer to the question. Please see the supplementary document relating to “Reporting Associations” for details.***

*Keys to past homeworks from quarters that I taught Biost 517 (e.g. HW #8 from 2012) or Biost 518 (e.g., HW #1 from 2014 or HWs #1, 3 from 2008) or Biost 536 (e.g. HW #3 from 2013) might be consulted for the presentation of inferential results. Note that the requirement to provide a paragraph describing your statistical methods was new last year, and thus keys prior to 2014 do not give explicit examples of a separate paragraph. However, many past keys provide this information as an introductory sentence.*

All questions relate to associations between death from any cause and serum C reactive protein (CRP) levels in a population of generally healthy elderly subjects in four U.S. communities. This homework uses the subset of information that was collected to examine inflammatory biomarkers and mortality. The data can be found on the class web page (follow the link to Datasets) in the file labeled inflamm.txt. Documentation is in the file inflamm.pdf. The data is in free-field format, and can be read into R by

read.table("http://www.emersonstatistics.com/datasets/inflamm.txt",header=T)

It can be read into Stata using the following code in a .do file.

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

Note that the first line of the text file contains the variable names, and will thus be converted to missing values. Similarly, there is some missing data recorded as ‘NA’, and those, too, will be converted to missing values. If you do not want to see all the warning messages, you can use the “quietly” prefix. You may want to go ahead and drop the first case using “drop in 1”, because it is just missing values.

Recommendations for risk of cardiovascular disease according to serum CRP levels are as follows (taken from the Mayo Clinic website):

|  |  |
| --- | --- |
| Below 1 mg/L | Low risk of heart disease |
| 1 - 3 mg/L | Average risk of heart disease |
| Above 3 mg/L | High risk of heart disease |

1. The observations of time to death in this data are subject to (right) censoring. Nevertheless, problems 2 – 6 ask you to dichotomize the time to death according to death within 4 years of study enrolment or death after 4 years. Why is this valid? Provide descriptive statistics that support your answer.

**Ans:**

**Methods:**

We created a dichotomized time-to-death variable indicating 4 years vital status dividing the total time in days by 365.25. If death was observed within 4 years of follow-up the dichotomized variable was set to 1 otherwise it was set to 0.

**Inference**:

It is valid to dichotomize time-to-death according to death within 4 years, because minimum time of follow-up among censored observations is 1480 days or about 4.05 years. Thus, the vital status of every individual is known at 4 years. We observed that 495 (9.9%) of the subjects died within 4 years (Table 1).

|  |  |
| --- | --- |
|  | N (%) |
| Died within 4 years | 495 (9.9) |
| Died after 4 years | 4505 (90.1) |

 Table 1. Proportion of persons who died before and after 4 years of follow-up.

1. Provide a suitable descriptive statistical analysis for selected variables in this dataset as might be presented in Table 1 of a manuscript exploring the association between serum CRP and 4 year all-cause mortality in the medical literature. In addition to the two variables of primary interest, you may restrict attention to age, sex, BMI, smoking history, cholesterol, and prior history of cardiovascular disease.

**Ans**:

**Methods:**

Blood C-Reactive Protein (CRP) level, a previously noted risk of cardiovascular disease, was categorized using low (< 1 mg/L), average (1-3 mg/L) and high (> 3 mg/L) cutoff values for risk of developing heart disease according to Mayo Clinic’s definition. Scatter plot of subjects’ age by CRP level was used prior to categorization to assess outliers. Subject’s age and sex at the study of enrollment, body mass index (BMI), smoking status, cholesterol, history of previous angina, MI, TIA or stroke, and death within 4 years of study enrollment were all described by CRP categories. Means, standard deviations, minimum and maximum values are presented for the continuous variables; age, BMI, and cholesterol. Percentages are presented for the categorical variables; sex, history of cardiovascular disease, smoking status, and death within 4 years of study enrollment. We considered the extreme CRP values credible and included them in our analyses as such (Figure 1). Subjects with missing CRP values were excluded. If subjects had missing risk factors information but had CRP data, they were kept in the analysis.

**Inference:**

Data on 5000 subjects was collected. Out of these, 4933 subjects had their serum CRP measured and 67 subjects had missing CRP values. After excluding the subjects with missing CRP, 6 subjects had missing smoking status information, 13 subjects had missing BMI measurement and 3 subjects had missing cholesterol measurement. Among 4933 subjects with available CRP measurement, 428 had CRP level less than 1 mg/L, 3330 had CRP level between 1 and 3 mg/L inclusive, and 1175 had CRP level greater than 3 mg/L. On average, males tend to be in the lower category of CRP. Subjects with low CRP level tend to be younger, smoke less, less likely to die within 4 years of follow-up, and have lower BMI on average, compared with subjects in the higher categories of CRP level. No such trend was observed for cholesterol (Table 1). Particularly, 4.91% of subjects died within 4 years of follow-up among subjects with CRP level less than 1 mg/L compared with 15.97% who died and had CRP level greater or equal to 3 mg/L. The total number of subjects who died within 4 years of follow up was 484 (9.82%).



Figure 1. Scatter plot of subjects age by CRP

|  |  |
| --- | --- |
|  | **Blood C-Reactive Protein Level (CRP)** |
|  | **< 1 mg/L(n=428)** | **1 - 3 mg/L(n=3330)** | **> 3 mg/L(n=1175)** | **Total(n=4933)** |
| Male (%) | 45.56% | 43.30% | 37.02% | 42.00% |
| Age (yrs)\* | 73.45 (5.80; 65.0 - 94.0) | 72.36 (5.52; 65.0 - 100) | 72.74 (5.58; 65.0 - 93.0) | 72.80 (5.56; 65.0 - 100) |
| BMI (kg/m2)\* | 23.82 (3.64; 15.6 - 38.6) | 26.39 (4.31; 14.7 - 53.2) | 28.45 (5.46; 15.3 - 58.8) | 26.66 (4.72; 14.7 - 58.8) |
| Cholesterol (mg/dl)\* | 206.00 (40.53; 109.0 - 407.0) | 212.83 (38.57; 73.0 - 363.0) | 210.50 (40.38; 97.0 - 430.0) | 211.68 (39.23; 73.0 - 430.0) |
| History of cardiovascular disease (%) | 18.22% | 21.47% | 28.77% | 22.93% |
| Smokers (%) | 9.60% | 11.01% | 16.43% | 12.18% |
| Death within 4 years (%) | 4.91% | 8.41% | 15.57% | 9.81% |

\* Mean (standard deviation; minimum – maximum)

Table 1. Descriptive statistics of study population by levels of serum CRP.

1. Perform a statistical analysis evaluating an association between serum CRP and 4 year all-cause mortality by comparing mean CRP values across groups defined by vital status at 4 years.

**Ans**:

**Methods**:

We compared the mean serum CRP between subjects who died within 4 years of follow-up and subjects who survived at least 4 years. A two samples t-test with unequal variance assumption was used to test the difference in the means. The 95% confidence interval for the difference in means was computed.

**Inference**:

The mean serum CRP for subjects who died within 4 years of follow-up was 5.38 mg/L (n=484), and for those survived at least 4 years of follow-up was 3.42 mg/L (n=4449). The estimated mean CRP of the subjects who died after 4 years was 1.95 mg/L higher than those who survived at least 4 years. The mean difference would not be unusual if the true mean difference was between 1.21 and 2.70 mg/L. Using the two sample t-test with unequal variance assumption, we rejected the null hypothesis that mean serum CRP is not different between subjects who died within 4 years and survived after 4 years of follow-up. The statistical evidence for the difference in means was sufficient at a 0.05 level of significance (two-sided p-value < 0.0001).

1. Perform a statistical analysis evaluating an association between serum CRP and 4 year all-cause mortality by comparing geometric mean CRP values across groups defined by vital status at 4 years. (Note that there are some measurements of CRP that are reported as zeroes. Make clear how you handle these measurements.)

**Ans**:

**Methods:**

The geometric mean was assessed after taking the log of the serum CRP values, while for subjects with 0 mg/L CRP measurement, their log serum CRP were replaced by a value of 0.5. This allowed measurements of CRP that are reported as zeroes to be included in the analysis (Figure 2). We compared the geometric mean serum CRP between subjects who died within 4 years and subjects who survived at least 4 years of follow-up. We used the two sample t-test with unequal variance assumption to test this mean difference. The 95% confidence interval for the difference in mean log CRP was computed. The exponents of the log means and confidence interval estimates were then reported.

**Inference:**

There were 428 subjects with 0 mg/L CRP measurement that were replaced by 0.5. The geometric mean serum CRP for subjects who died within 4 years of follow-up was 4.08 mg/L (n=484), and for those survived at least 4 years of follow-up was 2.26 mg/L (n=4449). The estimated geometric mean CRP of the subjects who died after 4 years was 0.72 mg/L higher than those who survived at least 4 years. The geometric mean difference would not be unusual if the true geometric mean difference was somewhere between 0.66 and 0.79 mg/L. Using the two sample t-test with unequal variance assumption, we rejected the null hypothesis that the geometric mean serum CRP is not different between subjects who died within 4 years and survived after 4 years of follow-up. The statistical evidence for the difference in the geometric means was sufficient at a 0.05 level of significance (two-sided p-value < 0.0001).



Figure 2 Distribution of log CRP for all subjects

1. Perform a statistical analysis evaluating an association between serum CRP and 4 year all-cause mortality by comparing the probability of death within 4 years across groups defined by whether the subjects have high serum CRP (“high” = CRP > 3 mg/L).

**Ans**:

**Methods:**

We assessed the probability of dying within 4 years of follow-up by comparing proportions between subjects with serum CRP level > 3 mg/L and ≤ 3 mg/L. Pearson chi-squared test for independence was used to test the difference in the proportions. The Wald test statistic was used to obtain the 95% confidence interval for the difference in probability of dying within 4 years of follow-up.

**Inference**:

There were 1175 (23.8%) subjects with serum CRP > 3 mg/L and 3758 (76.2%) subjects with CRP ≤ 3 mg/L. Subjects with missing CRP measurement (n=67) were excluded from the analysis. We observed that 8.01% of subjects with serum CRP ≤ 3 mg/L and 15.58% of subjects with serum CRP > 3 mg/L died within 4 years of follow-up. The estimated absolute probability of dying within 4 years of follow-up was 7.56% higher for subjects with serum CRP > 3 mg/L with a 95% confidence interval between 5.32% lower absolute probability 9.81% higher absolute probability. The statistical evidence for this difference in probability of dying within 4 years is sufficient (two-sided p-value < 0.0001). Thus, we reject the null-hypothesis that the survival probabilities are independent between subjects with serum CRP > 3 mg/L and ≤ 3 mg/L.

1. Perform a statistical analysis evaluating an association between serum CRP and 4 year all-cause mortality by comparing the odds of death within 4 years across groups defined by whether the subjects have high serum CRP (“high” = CRP > 3 mg/L).

**Methods**:

We compared the odds of dying within 4 years of follow-up between subjects with serum CRP > 3 mg/L and those with serum CRP ≤ 3 mg/L. Fisher’s exact test with two-sided p-value was computed to test if the odds ratio was different from 1, and the 95% confidence interval for the odds ratio was also computed.

**Inference**:

The odds of dying within 4 years of follow up was 0.087 for subjects with serum CRP ≤ 3 mg/L, and 0.18 for those with serum CRP > 3 mg/L. The observed odds ratio of dying comparing the two CRP groups was 2.12 with a 95% confidence interval between 1.74 and 2.58. The statistical evidence for the association between serum CRP and the odds of dying within 4 years is sufficient (two-sided p-value < 0.0001). Thus, we reject the null hypothesis that odds ratio for dying within 4 years is equal to 1.

1. Perform a statistical analysis evaluating an association between serum CRP and all-cause mortality over the entire period of observation of these subjects by comparing the instantaneous risk of death across groups defined by whether the subjects have high serum CRP (“high” = CRP > 3 mg/L).

**Methods**:

The survival distribution of subjects with CRP > 3 mg/L and subjects with CRP ≤ 3 was computed using Kaplan-Meier method. The difference in survival distribution over the entire period of observation was compared between the two groups using the Log-rank test. Comparing the instantaneous risk of death across the CRP groups, we used the Cox proportional hazards regression method to compute the hazard ratio and its 95% confidence interval.

**Inference**:

There were 1175 (23.8%) subjects with serum CRP > 3 mg/L and 3758 (76.2%) subjects with CRP ≤ 3 mg/L with their survival probability shown in Figure 1. Subjects with CRP > 3 mg/L tend to have an increase in survival probability than those with CRP ≤ 3 mg/L throughout the follow-up period. We estimate that the instantaneous risk of death was 68.72% higher for subjects with CRP ≤ 3 mg/L with a 95% confidence interval of (48.51%, 91.68%). Based on the Log-rank test, we reject the null hypothesis that the probability of survival over the entire period of follow-up is the same for subjects with CRP ≤ 3 mg/L and subjects with CRP > 3 mg/L (two-sided p-value < 0.0001).



Figure 3. Total survival probability distribution by CRP high vs low category

1. Supposing I had not been so redundant (in a scientifically inappropriate manner) and so prescriptive about methods of detecting an association, what analysis would you have preferred *a priori* in order to answer the question about an association between mortality and serum CRP? Why?

The best analysis I prefer is to use the low (< 1 mg/L), average (1-3 mg/L) and high (> 3 mg/L) CRP cutoff values for risk of developing heart disease according to Mayo Clinic’s definition and looking at the association between this variable and all-cause mortality over the entire period. Categorizing CRP into three groups increases the sensitivity of our study, is scientifically feasible, and allows us to compare the survival distribution of a reference CRP category and the other two categories. In our data, we attained sufficient sample size in each CRP category. I preferred all-cause mortality over the entire period of observation because only 9.9% died within 4 years of follow up, while 22.4% died over the entire period. Thus, the latter would be more statistically precise. For the method I chose above, the data is censored so I would use the Kaplan Meir method to determine the association between mortality and serum CRP three-groups.

**APPENDIX**

STATA and R codes for problem set 1:

PROBLEM 1:

quietly: 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

list in 1/5, table

/\*Summarize time to death in days for censored individuals \*/

summ ttodth if death==0

/\*

 Variable | Obs Mean Std. Dev. Min Max

-------------+--------------------------------------------------------

 ttodth | 3879 2603.711 413.5922 1480 2942

\*/

/\*

Minimum time of follow-up among censored observations is 1480 days or about 4 years

Thus, the vital status of every individual is known at 4 years.

 Now, we create a dichotomized time-to-death, according to death within 4 years of study enrollment

 or death after 4 years.

\*/

drop deadin4

g deadin4= 0

replace deadin4= 1 if ttodth <= 4 \* 365.25

tabulate deadin4

/\*

 deadin4 | Freq. Percent Cum.

------------+-----------------------------------

 0 | 4,505 90.10 90.10

 1 | 495 9.90 100.00

------------+-----------------------------------

 Total | 5,000 100.00

\*/

/\*Recommendations for risk of cardiovascular disease according to

serum CRP levels are as follows (taken from the Mayo Clinic website):

\*/

/\*Convert to numeric \*/

destring crp, replace

drop crpct

recode crp 4/max=3 1/3=2 min/1=1, gen(crpct)

tab crpct

/\*

 RECODE of |

 crp (crp) | Freq. Percent Cum.

------------+-----------------------------------

 1 | 428 8.68 8.68

 2 | 2,629 53.29 61.97

 3 | 1,876 38.03 100.00

------------+-----------------------------------

 Total | 4,933 100.00

\*/

PROBLEM 2:

/\*

2. Provide a suitable descriptive statistical analysis for selected

variables in this dataset as might be presented in Table 1 of a manuscript

exploring the association between serum CRP and 4 year all-cause mortality

in the medical literature. In addition to the two variables of primary interest,

you may restrict attention to age, sex, BMI, smoking history, cholesterol,

and prior history of cardiovascular disease.

\*/

destring age male bmi smoker cholest prevdis, replace

tabstat age male bmi smoker cholest prevdis, by (crpct) stat(n mean sd min q max) col(stat) long

tabstat deadin4, by(crpct) stat(n mean sd min q max) col(stat) long

tabstat age male bmi smoker cholest prevdis crp, by (deadin4) stat(n mean sd min q max) col(stat) long

tabulate crpct male, row

tabulate crpct smoker, row

tabulate crpct prevdis, row

tabulate crpct deadin4, row

PROBLEM 3:

g logcrp2 = 0.5 if crp=0

replace logcrp2=log(crp) if crp>0

ttest crp, by(deadin4) unequal

/\*

Two-sample t test with unequal variances

------------------------------------------------------------------------------

 Group | Obs Mean Std. Err. Std. Dev. [95% Conf. Interval]

---------+--------------------------------------------------------------------

 0 | 4449 3.422117 .0880345 5.871972 3.249526 3.594709

 1 | 484 5.376033 .3680768 8.097691 4.652803 6.099263

---------+--------------------------------------------------------------------

combined | 4933 3.613825 .0876014 6.152715 3.442087 3.785563

---------+--------------------------------------------------------------------

 diff | -1.953916 .3784582 -2.697348 -1.210484

------------------------------------------------------------------------------

 diff = mean(0) - mean(1) t = -5.1628

Ho: diff = 0 Satterthwaite's degrees of freedom = 539.648

 Ha: diff < 0 Ha: diff != 0 Ha: diff > 0

 Pr(T < t) = 0.0000 Pr(|T| > |t|) = 0.0000 Pr(T > t) = 1.0000

\*/

PROBLEM 4:

/\*

T-test in SAS for log transformed CRP

 The TTEST Procedure

 Variable: lcrp2

 deadin4 N Mean Std Dev Std Err Minimum Maximum

 0 4449 0.8170 0.8203 0.0123 0 4.6821

 1 484 1.1407 0.9521 0.0433 0 4.0073

 Diff (1-2) -0.3237 0.8342 0.0399

 deadin4 Method Mean 95% CL Mean Std Dev 95% CL Std Dev

 0 0.8170 0.7929 0.8411 0.8203 0.8037 0.8378

 1 1.1407 1.0556 1.2257 0.9521 0.8957 1.0162

 Diff (1-2) Pooled -0.3237 -0.4020 -0.2454 0.8342 0.8180 0.8510

 Diff (1-2) Satterthwaite -0.3237 -0.4121 -0.2353

 Method Variances DF t Value Pr > |t|

 Pooled Equal 4931 -8.11 <.0001

 Satterthwaite Unequal 563.77 -7.19 <.0001

 Equality of Variances

 Method Num DF Den DF F Value Pr > F

 Folded F 483 4448 1.35 <.0001

\*/

di exp(0.817), exp(1.407), exp(-.3237), exp(-.4121), exp(-.2353)

PROBLEM 5

/\*

5. Perform a statistical analysis evaluating an association

between serum CRP and 4 year all-cause mortality by comparing

the probability of death within 4 years across groups defined by

whether the subjects have high serum CRP (“high” = CRP > 3 mg/L).

\*/

/\*Exclude the missing values \*/

drop crpgt3

generate crpgt3=.

replace crpgt3=1 if crp > 3 & crp != .

replace crpgt3=0 if crp <= 3 & crp != .

tab crpgt3

tab crpct

/\*

SAS:

 crpgt3 Frequency Percent Frequency Percent

 ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ

 0 3758 76.18 3758 76.18

 1 1175 23.82 4933 100.00

STATA:

 crpgt3 | Freq. Percent Cum.

------------+-----------------------------------

 0 | 3,758 76.18 76.18

 1 | 1,175 23.82 100.00

------------+-----------------------------------

 Total | 4,933 100.00

\*/

/\* Convert time to death from days to months and from days to years\*/

g ttodthm = ttodth/30.4

g ttodthy = ttodth/365.25

summ ttodthm ttodthy ttodth

/\*

 Variable | Obs Mean Std. Dev. Min Max

-------------+--------------------------------------------------------

 ttodthm | 5000 77.90728 22.26591 .1644737 96.77631

 ttodthy | 5000 6.484275 1.853207 .0136893 8.054757

 ttodth | 5000 2368.381 676.8838 5 2942

\*/

univar ttodthm, by (crpgt3)

cs deadin4 crpgt3, or

/\*

 | crpgt3 |

 | Exposed Unexposed | Total

-----------------+------------------------+------------

 Cases | 183 301 | 484

 Noncases | 992 3457 | 4449

-----------------+------------------------+------------

 Total | 1175 3758 | 4933

 | |

 Risk | .1557447 .0800958 | .0981147

 | |

 | Point estimate | [95% Conf. Interval]

 |------------------------+------------------------

 Risk difference | .0756489 | .0531723 .0981254

 Risk ratio | 1.94448 | 1.637791 2.308599

 Attr. frac. ex. | .4857237 | .3894216 .5668368

 Attr. frac. pop | .1836517 |

 Odds ratio | 2.118714 | 1.740572 2.579029 (Cornf

> ield)

 +-------------------------------------------------

 chi2(1) = 57.89 Pr>chi2 = 0.0000

\*/

/\*To obtain a p-value for OR using Fischer's exact test with 2-sided p-value \*/

cc deadin4 crpgt3, exact

/\*

cc deadin4 crpgt3, exact

 Proportion

 | Exposed Unexposed | Total Exposed

-----------------+------------------------+------------------------

 Cases | 183 301 | 484 0.3781

 Controls | 992 3457 | 4449 0.2230

-----------------+------------------------+------------------------

 Total | 1175 3758 | 4933 0.2382

 | |

 | Point estimate | [95% Conf. Interval]

 |------------------------+------------------------

 Odds ratio | 2.118714 | 1.729676 2.591835 (exact

> )

 Attr. frac. ex. | .5280155 | .421857 .614173 (exact

> )

 Attr. frac. pop | .1996422 |

 +-------------------------------------------------

 1-sided Fisher's exact P = 0.0000

 2-sided Fisher's exact P = 0.0000

\*/

PROBLEM 6 and PROBLEM 7:

stset ttodthm death

sts graph, by(crpgt3) ylabel(0.70(0.05)1.00) plot1opts(lcolor(black)) plot2opts(lcolor(blue))

stcox crpgt3, robust

/\*

 failure \_d: death

 analysis time \_t: ttodthm

Iteration 0: log pseudolikelihood = -9175.9083

Iteration 1: log pseudolikelihood = -9146.2154

Iteration 2: log pseudolikelihood = -9145.5241

Iteration 3: log pseudolikelihood = -9145.524

Refining estimates:

Iteration 0: log pseudolikelihood = -9145.524

Cox regression -- Breslow method for ties

No. of subjects = 4933 Number of obs = 4933

No. of failures = 1109

Time at risk = 385305.1317

 Wald chi2(1) = 64.56

Log pseudolikelihood = -9145.524 Prob > chi2 = 0.0000

------------------------------------------------------------------------------

 | Robust

 \_t | Haz. Ratio Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

 crpgt3 | 1.687192 .1098327 8.04 0.000 1.485091 1.916796

------------------------------------------------------------------------------

HR = 1.687

\*/

PLOTS IN R

d1 <- read.table("http://www.emersonstatistics.com/datasets/inflamm.txt",header=T)

d1[1:4,]

d2<-transform(d1, ttodth2=as.numeric(ttodth))

summary(d2$ttodth)

summary(d2$crp)

d2$crpgt3<-ifelse(d2$crp>3,1,0)

table(d2$crpgt3)

d2$ttodthm<- d1$ttodth/30.4

summary(d2$ttodthm)

d2$logcrp <- ifelse(d2$crp==0,0.5,log(d2$crp))

hist(d2$logcrp,col="grey",main=" ",xlab="Log CRP")

hist(d2$crp,col="grey",main=" ",xlab="CRP in mg/L")

plot(jitter(d2$age) ~ jitter(d2$crp), pch=20, xlab="CRP in md/L", ylab="Age in Years")

plot(jitter(d2$age) ~ jitter(d2$logcrp,1), pch=20, xlab="Log CRP", ylab="Age in Years")

plot(jitter(d2$age) ~ d2$logcrp, pch=20, xlab="Log CRP", ylab="Age in Years")

summary(d2$logcrp)

summary(d2$crp)

install.packages("survfit")

library(survival)

plot(survfit(formula = Surv(ttodthm, death)~ crpgt3, data = d2, conf.type="none"),

 lty=0, xlab="Months", ylab="Survival Probability" )

timestrata.surv <- survfit( Surv(ttodthm, death)~ strata(crpgt3), d2, conf.type="log-log")

plot(timestrata.surv, lty=c(1,3), ylim=c(0.6,1), mark.time=FALSE, xlab="Time in Months", ylab="Survival Probability")

 legend(50, 1.0, c("CRP <= 3 mg/L", "CRP > 3 mg/L") , lty=c(1,3) )