Biost 517: Applied Biostatistics I
Emerson, Fall 2010

Homework #5 Key
November 3, 2010

Written problems: To be handed in at the beginning of class on Wednesday, November 3, 2010.

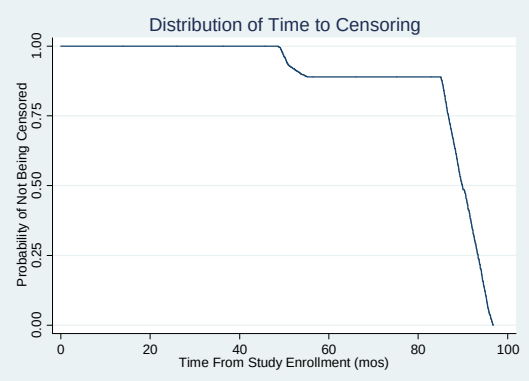
Questions for Biost 514 and Biost 517:

The following problems make use of a dataset exploring the prognostic value of certain biomarkers of inflammation on all cause mortality. The documentation file inflamm.doc and the data file inflamm.txt can be found on the class web pages.

1. In studies with censored observations of time to some event, our ability to answer specific scientific questions will often depend upon the distribution of censoring times. That is, we need to understand the times that we followed each patient. However, we only have partial information on this distribution. For instance, if we are ultimately investigating patient survival, we may want to understand how long we followed the patients: Was it 3 years, 30 years, 300 years? Was it 3 years for some patients and 300 years for others? When patients' survival times are censored, we know exactly the limits of our follow-up. But for patients who died, we do not know when we might have lost those patients to further follow-up. Luckily the Kaplan-Meier estimator comes to our rescue in this situation. By creating an indicator of censoring (0= not censored, 1= censored), we can use the KM estimates to describe the pattern of censoring.
    a. Provide suitable statistics for the distribution of times to censoring for observations of death. In particular, consider whether you can estimate the minimum time of follow-up for these patients.

Answer:

The following figure is a plot of the Kaplan-Meier estimates of the censoring time distribution. *(Note that the actual accrual to this study occurred over two periods. Most subjects were accrued over about a one year period, and then approximately two years later, a second cohort of minority enriched participants were recruited. All subjects were then followed from the time of their accrual.)*



The following table presents descriptive statistics for the censoring distribution including the cumulative probability of remaining uncensored at yearly intervals, the time at which 75%, 50%, and 25% of the sample remained uncensored, and the mean censoring time. Because the minimal follow-up time corresponds to a death rather than an censored observation, we cannot be sure of

the time that an individual might have been censored. However the Kaplan-Meier estimate of the censoring distribution suggests that all subjects (100%) would have been followed at least 4 years.

	Censoring of Mortality	**Censoring of CV Mortality**
Probability Uncensored		
12 month	1.000	0.992
24 month	1.000	0.979
36 month	1.000	0.964
48 month	1.000	0.946
60 month	0.889	0.820
72 month	0.889	0.801
84 month	0.889	0.777
96 month	0.038	0.032
Time with Percentage Uncensored (nonths)		
75% uncensored	86.7	85.5
50% uncencored	89.9	88.8
25% uncensored	93.4	92.8
Mean Time of Follow-up (months)	86.4	81.6

b. Suppose we want to divide individual patients into groups who die within 3 years and those who do not. On the basis of your answer to part a, will we be able to do so? What about groups of patients defined by 5 year survival—can we do that?

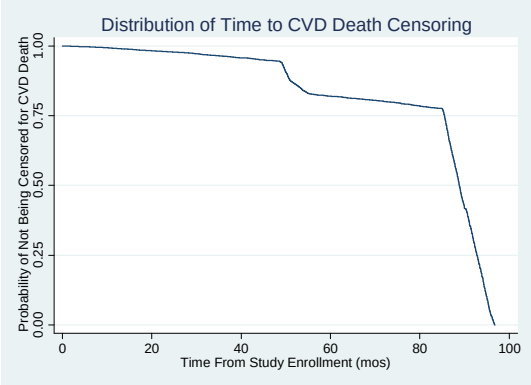
Answer:

Because the earliest censored observation was after more than four years of follow-up, we can certainly identify which individual subjects died within 3 years and which survived for more than 3 years. However, many observations are censored between 4 years and 5 years of follow-up. Any of those censored observations could have died prior to 5 years, thus we cannot divide the sample into subjects who did or did not die within 5 years. (Kaplan-Meier estimates allow us to guess the proportion who died within 5 years, but not necessarily which subjects they were.)

c. Provide suitable statistics for the distribution of times to censoring for observations of cardiovascular death. Again, consider whether you can estimate the minimum time of follow-up for these patients.

Answer:

The following figure is a plot of the Kaplan-Meier estimates of the censoring time distribution for censoring of cardiovascular deaths. Note that subjects who died of other causes had censored observations for the time they would have died from cardiovascular disease (as opposed to say, drowning, car accidents, cancer, being smashed by a meteor). Tabled descriptive statistics are presented in the table given above. There were definitely some subjects whose CV death time was censored early by the competing risk of other cause mortality.



d. Suppose we want to divide individual patients into groups who die from cardiovascular disease within 3 years and those who do not. On the basis of your answer to part c, will we be able to do so? What about groups of patients defined by 5 year mortality from cardiovascular disease—can we do that?

Answer:

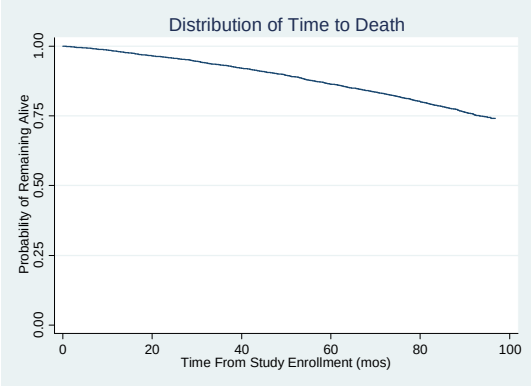
Because the earliest censored observation was after only about a month of follow-up, we cannot identify all individual subjects who died (or would have died) within 3 years of cardiovascular causes and which would have survived cardiovascular death for more than 3 years. Nor can we identify all such subjects over a 5 year period. (Again, Kaplan-Meier estimates allow us to guess the proportion who would have died of cardiovascular causes within 3 or 5 years if propensity for cardiovascular death was independent of other causes of mortality, but not necessarily which subjects they were.)

2. We are interested in estimating the probability of a patient dying from any cause in the years following accrual to the study.

   a. Provide suitable descriptive statistics for the distribution of times to death from any cause for all patients in the study.

Answer:

The following figure is a plot of the Kaplan-Meier estimates of the distribution of time to death.



The following table presents descriptive statistics including the probability of survival beyond yearly intervals, the times at which 90%, 80%, and 75% or the patients are estimated to still be alive, and the mean months alive during the first 96.8 months of follow-up.

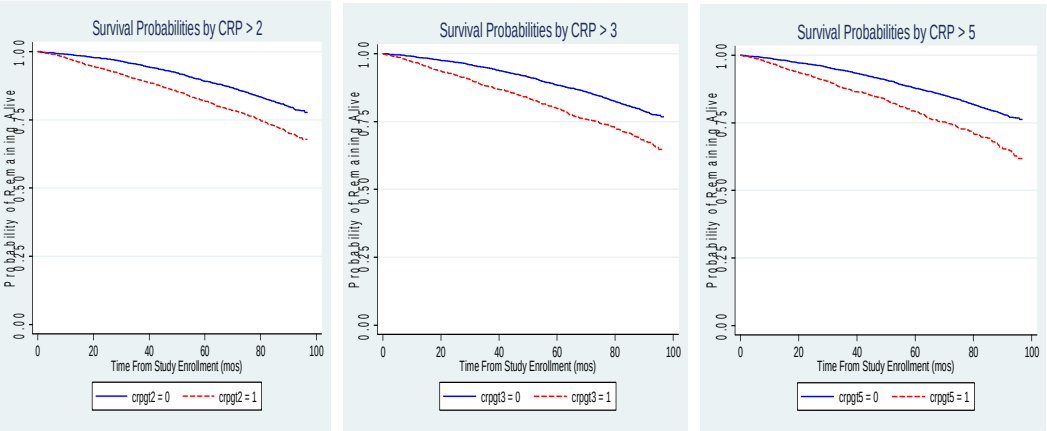
	Overall	CRP < 2	CRP > 2	CRP < 3	CRP > 3	CRP < 5	CRP > 5
Sample Size	5000	3057	1876	3758	1175	4135	798
Survival Probabilities							
12 month	0.982	0.989	0.972	0.988	0.967	0.986	0.967
24 month	0.959	0.975	0.936	0.971	0.926	0.967	0.926
36 month	0.931	0.953	0.898	0.948	0.881	0.943	0.877
48 month	0.901	0.927	0.861	0.920	0.844	0.913	0.843
60 month	0.864	0.891	0.820	0.884	0.800	0.878	0.793
72 month	0.829	0.860	0.780	0.853	0.755	0.845	0.747
84 month	0.787	0.820	0.731	0.810	0.710	0.804	0.696
96 month	0.741	0.777	0.679	0.769	0.647	0.763	0.619
Time with Specified Percentage Survival (nonths)							
90% survival	48.3	57.6	35.1	54.5	30.9	53.1	30.7
80% survival	80.2	90.0	64.6	87.1	59.6	85.2	58.2
75% survival	93.4	NA	79.9	NA	73.8	NA	70.9
Restricted Mean Survival (months)							
96.8 month	86.1	88.3	82.5	87.7	80.9	87.2	80.3

b. Produce a plot of survival curves stratified by the groups defined by whether the C-reactive protein (CRP) value was higher than 2 mg/l or not. Produce a table of estimates of the 90th, 80th, and 75th percentiles of the survival distribution by CRP strata. Also include in that table the estimated probabilities of surviving for 3, 5, and 8 years for each stratum. Are the estimates suggestive that CRP level is associated with mortality? Give descriptive statistics supporting your answer.

Answer:

The following figures plot of the Kaplan-Meier estimates of the distribution of time to death within strata defined by whether of not CRP > 2 (left panel), whether or not CRP > 3 (middle panel), and whether or not CRP > 5 (right panel). In each case, the stratum with the higher CRP measurements tend to show worse survival, as seen by the vertical separation between the curves at each time point. This is also seen in the above table, which also presents the numerical estimates of survival at yearly intervals, as well as providing the time points at which 90%, 80%, or 75% of subjects are estimated to still be alive within each stratum. The table also includes the average number of months alive during the first 96.8 months following study accrual.

These analyses are suggestive of an association between CRP level and time to death: For instance, the 5 year survival probability is estimated to be 89.1% among those subjects with CRP less than 2 ng/ml, but only estimated to be 82.0% amongh subjects with CRP greater than 2 ng/ml. Similar separation of the curves are seen at other time points and with the other thresholds for dichotomizing the CRP values.



c. Repeat part b using thresholds of 3 mg/l and 5 mg/l for CRP.

Answer:

(I answered both parts b and c above.)

3. Suppose we are interested in using the CRP to predict whether a patient will still be alive three years after study accrual.
   a. In our sample, what is the prevalence of death within 3 years?

Answer:

From the table given in problem 2a, we see that we estimate 93.1% survival at 3 years, thus the proportion that are dead is estimated to be 6.9%.

b. In our sample, what is the prevalence of a CRP greater than 2 mg/l?

Answer:

From the table given in problem 2a, we see that 1876 subjects were known to have CRP greater than 2 ng/ml, and 3057 were known tohave CRP below that threshold. Hence, assuming that CRP values are missing completely at random on the 67 subjects with missing values would allow us to estimate a prevalence of 38.0% of subjects with CRP greater than 2 ng/ml.

c. Suppose we consider a CRP greater than 2 mg/l to be a "positive" test result. What are the sensitivity and specificity of such a diagnostic criterion? Briefly explain how these were calculated.

Answer:

Because our data represent a cross-sectional sample of 5,000 subjects (that is, we restricted neither the proportion of deaths within 3 years nor the proportion with high CRP), and because we observed all subjects for at least 3 years, we can merely produce a crosstabulation of subjects according to death within 3 years or not and according to CRP greater than 2 ng/ml or not. Such a crosstabulation results in an observation that 191 "positive" tests were observed among the 334 subjects dying within 3 years, yielding a sensitivity of 57.2%. Similarly, 2,914 "negative" tests were observed among the 4,599 subjects surviving at least 3 years, yielding a specificity of 63.4%

d. If the sample accurately reflects the patient population of interest, what are the positive and negative predictive values of such a diagnostic criterion? Briefly explain how these were calculated.

Answer:

Because our data represent a cross-sectional sample of 5,000 subjects (that is, we restricted neither the proportion of deaths within 3 years nor the proportion with high CRP), and because we observed all subjects for at least 3 years, we can merely produce a crosstabulation of subjects according to death within 3 years or not and according to CRP greater than 2 ng/ml or not. Such a crosstabulation results in an observation that 191 deaths within 3 years were observed among the 1,876 subjects with a "positive" test, yielding a positive predictive value of 10.2%. Similarly, 2,914 subjects surviving 3 years were observed among the 3,057 subjects with "negative" test results, yielding a predictive value of the negative of 95.3%

e. Repeat parts b, c, and d using thresholds of 3 mg/l and 5 mg/l. (You need not explain how they were calculated, just include the sensitivity, specificity, predictive value of a positive, and predictive value of a negative in a table.)

Answer:

When "disease" is defined as death within 3 years, the following table presents the prevalence of a "positive" test, its sensitivity, its specificity, its positive predictive value, and its negative predictive value as a function of the threshold used to declare positivity. It can be seen that as we increase the threshold, we increase specificity and decrease sensitivity. Similarly, we increase the predictive value of the positive, but decrease the predictive value of the negative. (A more complete analysis might have considered every possible threshold for declaring test "positivity". A graph comparing the sensitivity to 1 minus the specificity as we vary that threshold is called a "Receiver Operating Characteristic Curve" (or just ROC curve- the name comes from engineering.)

	Threshold for Positivity		
	CRP > 2 ng/ml	**CRP > 3 ng/ml**	**CRP > 5 ng/ml**
Proportion Positive	**38.0%**	**23.8%**	**16.2%**
Sensitivity	**57.2%**	**41.9%**	**29.3%**
Specificity	**63.4%**	**77.5%**	**84.8%**
Pred Val Pos	**10.2%**	**11.9%**	**12.3%**
Pred Val Neg	**95.3%**	**94.8%**	**94.3%**

4. Now suppose we are interested in using the CRP to predict whether a patient will still be alive five years after study accrual.

a. In our sample, what is the estimated prevalence of death within 5 years?

Answer:

From the table given in problem 2a, we see that we estimate 86.4% survival at 5 years, thus the proportion that are dead is estimated to be 13.6%.

b. Suppose we consider a CRP greater than 2 mg/l to be a "positive" test result. Can you calculate the sensitivity and specificity of such a diagnostic criterion? If so, do so. If not, briefly explain why not.

Answer:

We cannot do this as easily as we did it in problem 3, because we have censored observations before the desired five year period of observation is over. However, if we can obtain the predictive value of the positive and the predictive value of the negative, we can use Bayes rule along with the probability of a positive test to obtain these values. I will therefore revisit this question after answering part c.

c. If the sample accurately reflects the patient population of interest, can you calculate the positive and negative predictive values of such a diagnostic criterion? If so, do so. If not, briefly explain why not.

Answer:

We cannot do this as easily as we did it in problem 3, because we have censored observations before the desired five year period of observation is over. However, it is not very hard to use Kaplan-Meier estimates to obtain the desired quantities. In fact, they were partially given in the table from problem 2a.

- **Among subjects with CRP < 2 ng/ml, 89.1% are estimated to survive 5 years. This is the predictive value of the negative.**
- **Among subjects with CRP > 2 ng/ml, 82.0% are estimated to survive 5 years. Hence 18.0% are estimated to die within 5 years, and this is the predictive value of the positive.**

Now we can use Bayes rule to "reverse the conditional probabilities". (I actually used Excel with output from Stata.)

$$Sensitivity = \Pr(Pos \mid Disease) = \frac{\Pr(Disease \mid Pos) \times \Pr(Pos)}{\Pr(Disease \mid Pos) \times \Pr(Pos) + \Pr(Disease \mid Neg) \times \Pr(Neg)}$$

$$= \frac{0.180 \times 0.380}{0.180 \times 0.380 + 0.109 \times 0.620} = 0.504$$

$$Specificity = \Pr(Neg \mid Health) = \frac{\Pr(Health \mid Neg) \times \Pr(Neg)}{\Pr(Health \mid Neg) \times \Pr(Neg) + \Pr(Health \mid Pos) \times \Pr(Pos)}$$

$$= \frac{0.891 \times 0.620}{0.891 \times 0.620 + 0.820 \times 0.380} = 0.639$$

d. Repeat parts b and c using thresholds of 3 mg/l and 5 mg/l. (You need not explain how they were calculated, just include the sensitivity, specificity, predictive value of a positive, and predictive value of a negative in a table.)

Answer:

The following table presents the prevalence of a "positive" test, its sensitivity, its specificity, its positive predictive value, and its negative predictive value as a function of the threshold used to declare positivity and with a definition of "disease" based on death within 5 years. It can be seen that as we increase the threshold, we increase specificity and decrease sensitivity. Similarly, we increase the predictive value of the positive, but decrease the predictive value of the negative. (A more complete analysis might have considered every possible threshold for declaring test "positivity". A graph comparing the sensitivity to 1 minus the specificity as we vary that threshold is called a "Receiver Operating Characteristic Curve" (or just ROC curve- the name comes from engineering.)

	Threshold for Positivity		
	CRP > 2 ng/ml	**CRP > 3 ng/ml**	**CRP > 5 ng/ml**
Proportion Positive	**38.0%**	**23.8%**	**16.2%**
Sensitivity	**50.4%**	**35.1%**	**24.6%**
Specificity	**63.9%**	**78.0%**	**85.2%**
Pred Val Pos	**18.0%**	**20.0%**	**20.7%**
Pred Val Neg	**89.1%**	**88.4%**	**87.8%**

5. Suppose instead that the sample that we obtained undersampled patients who would actually die.
   a. If the true prevalence of death within three years in the target population were 20%, what would be the positive and negative predictive values of the diagnostic criterion based on a CRP greater than 2 ng/ml for predicting death within three years? Briefly explain how these were calculated.

Answer:

We now use Bayes rule to compute the positive and negative predictive values from the sensitivity and specificity obtained in problem 3 and the new prevalence.

$$PV+ = \Pr(Disease \mid Pos) = \frac{\Pr(Pos \mid Disease)\times \Pr(Disease)}{\Pr(Pos \mid Disease)\times \Pr(Disease) + \Pr(Pos \mid Health)\times \Pr(Health)}$$

$$= \frac{Sens \times \Pr ev}{Sens \times \Pr ev + (1 - Spec) \times (1 - \Pr ev)}$$

$$= \frac{0.572 \times 0.200}{0.572 \times 0.200 + 0.366 \times 0.800} = 0.281$$

$$PV- = \Pr(Health \mid Neg) = \frac{\Pr(Neg \mid Health)\times \Pr(Health)}{\Pr(Neg \mid Health)\times \Pr(Health) + \Pr(Neg \mid Disease)\times \Pr(Disease)}$$

$$= \frac{Spec \times (1 - \Pr ev)}{Spec \times (1 - \Pr ev) + (1 - Sens) \times \Pr ev}$$

$$= \frac{0.634 \times 0.800}{0.634 \times 0.800 + 0.428 \times 0.200} = 0.856$$

   b. Repeat part a using a threshold of a CRP greater than 3 mg/l and 5 mg/l.

Answer:

The following table presents the prevalence of a "positive" test, its sensitivity, its specificity, its positive predictive value, and its negative predictive value as a function of the threshold used to declare positivity and with a definition of "disease" based on death within 3 years.

	Threshold for Positivity		
	CRP > 2 ng/ml	**CRP > 3 ng/ml**	**CRP > 5 ng/ml**
Prevalence of Disease	**20.0%**	**20.0%**	**20.0%**
Sensitivity	**57.2%**	**41.9%**	**29.3%**
Specificity	**63.4%**	**77.5%**	**84.8%**
Pred Val Pos	**28.1%**	**31.8%**	**32.5%**

Pred Val Neg	**85.6%**	**84.2%**	**82.8%**

c. If the true prevalence of death within five years in the target population were 20%, can you estimate the positive and negative predictive values of the diagnostic criterion based on a CRP greater than 2 ng/ml for predicting death within 5 years? If so, do so. If not, briefly explain why not.

Answer:

We use the exact same methods: Bayes rule, but with the Sensitivity and Specificity as computed in problem 4c-d. The following table presents the prevalence of a "positive" test, its sensitivity, its specificity, its positive predictive value, and its negative predictive value as a function of the threshold used to declare positivity and with a definition of "disease" based on death within 5 years.

	Threshold for Positivity		
	CRP > 2 ng/ml	**CRP > 3 ng/ml**	**CRP > 5 ng/ml**
Prevalence of Disease	**20.0%**	**20.0%**	**20.0%**
Sensitivity	**50.4%**	**35.1%**	**24.6%**
Specificity	**63.9%**	**78.0%**	**85.2%**
Pred Val Pos	**25.9%**	**28.5%**	**29.3%**
Pred Val Neg	**83.8%**	**82.8%**	**81.9%**