Biost 517: Applied Biostatistics I
Emerson, Fall 2012

Homework #5 Key
October 28, 2012

Written problems: To be handed in at the beginning of class on Friday, November 2, 2012.

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. Some subjects are missing data for fibrinogen, but for the purposes of this homework we will presume that such data is missing completely at random (MCAR).

1. Suppose we are interested in using the fibrinogen levels (a marker of inflammation) 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:

In this data, observation of time to death was subject to censoring. However, no subject was censored prior to 4.05 years. So while we could use Kaplan-Meier estimates to estimate the prevalence of death within 3 years, the estimate would just be the same thing that we would get if we create a new variable indicating death within 3 years, a new variable indicating a fibrinogen level above 400 mg/dl, and then create the 2x2 "contingency table".

	3 Year Vital Status		Total
Fibrinogen	**Alive**	**Dead**	
≤ 400 mg/dl	**4,101**	**251**	**4,352**
> 400 mg/dl	**485**	**78**	**563**
Total	**4,596**	**329**	**4,915**

From the table given above, we see that 329 of 4,915 patients with available fibrinogen levels were dead within 3 years from the time of study accrual. Thus the incidence of death within 3 years ("prevalence of disease") is estimated to be 6.69%.

(Note that I used the MCAR assumption to decide that I could just omit all cases having missing data for fibrinogen. Under that same assumption, however, the probability that would be estimated using all 5,000 cases would also be valid and slightly more precise. I took the approach I did in order to highlight the use of the single contingency table that could be used in this cross-sectional sampling of the population, with longitudinal follow-up for mortality.)

    b. In our sample, what is the prevalence of a fibrinogen greater than 400 mg/dl?

Answer:

From the table given above, we see that 563 subjects were known to have fibrinogen greater than 400 mg/dl, and 4,352 were known to have fibrinogen below that threshold. Hence, assuming that fibrinogen values are missing completely at random on the 85 subjects with

missing values would allow us to estimate a prevalence of 11.5% of subjects with fibrinogen greater than 400 mg/dl.

c. Suppose we consider a fibrinogen greater than 400 mg/dl 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 fibrinogen), because we presume any subjects missing fibrinogen levels are MCAR, 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 fibrinogen greater than 400 mg/dl or not. Such a crosstabulation (as given in the table above) results in an observation that 78 "positive" tests were observed among the 329 subjects dying within 3 years, yielding a sensitivity of 23.7%. Similarly, 4,101 "negative" tests were observed among the 4,586 subjects surviving at least 3 years, yielding a specificity of 89.4%

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 fibrinogen), because we presume any subjects missing fibrinogen levels are MCAR, 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 fibrinogen greater than 400 mg/dl or not. Such a crosstabulation (as given in the table above) results in an observation that 78 deaths within 3 years were observed among the 563 subjects with a "positive" test, yielding a positive predictive value of 13.9%. Similarly, 4,101 subjects surviving 3 years were observed among the 4,352 subjects with "negative" test results, yielding a predictive value of the negative of 94.2%

e. Suppose instead that the sample that we obtained undersampled patients who would actually die. If the true prevalence of death within three years in the target population were 50%, what would be the positive and negative predictive values of the diagnostic criterion based on a fibrinogen greater than 400 mg/l 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 1c and the new prevalence.

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

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

$$= \frac{0.237 \times 0.500}{0.237 \times 0.500 + 0.106 \times 0.500} = 0.691$$

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

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

$$= \frac{0.894 \times 0.500}{0.894 \times 0.500 + 0.763 \times 0.500} = 0.540$$

f. Repeat parts b, c, and d using thresholds of 350 mg/dl. (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.

	Fibrinogen > 350 mg/dl	**Fibrinogen > 400 mg/dl**
Proportion Positive	**30.3%**	**11.5%**
Sensitivity	**45.6%**	**23.7%**
Specificity	**70.8%**	**89.4%**
Pred Val Pos	**10.1%**	**13.9%**
Pred Val Neg	**94.8%**	**94.2%**

2. Now suppose we are interested in using the fibrinogen 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:

Because the earliest censoring time is at 4.05 years, we need to use Kaplan-Meier estimates to estimate this "prevalence of disease". From such an analysis using all subjects (whether or not they have available fibrinogen levels) we estimate 86.3% survival at 5 years, thus the proportion that are dead is estimated to be 13.7%.

(Note that the missing completely at random assumption would argue that using all the data or using only subjects with available fibrinogen levels should give asymptotically consistent estimates of the desired probability. Because I cannot use contingency table analyses on this problem, I figured that I might as well use all available data for each part.)

b. If the sample accurately reflects the patient population of interest, can you calculate the positive and negative predictive values of a diagnostic criterion based on a fibrinogen greater than 400 mg/dl? 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. All we have to do is obtain KM estimates of 5 year survival within strata defined by fibrinogen above or below 400 mg/dl.

- **Among subjects with fibrinogen $\leq$ 400 mg/dl, 87.7% are estimated to survive 5 years. This is the predictive value of the negative.**
- **Among subjects with fibrinogen > 400 mg/dl, 76.8% are estimated to survive 5 years. Hence 23.2% are estimated to die within 5 years, and this is the predictive value of the positive.**

c. (Biost 514 or bonus for Biost 517): Suppose we consider a fibrinogen greater than 400 mg/dl 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.

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

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

$$= \frac{0.232 \times 0.115}{0.232 \times 0.115 + 0.123 \times 0.885} = 0.197$$

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

$$= \frac{0.877 \times 0.885}{0.877 \times 0.885 + 0.768 \times 0.115} = 0.898$$

d. (Biost 514 or bonus for Biost 517): If the true prevalence of death within 5 years in the target population were 50%, 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 now use Bayes rule to compute the positive and negative predictive values from the sensitivity and specificity obtained in problem 2c and the new prevalence.

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

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

$$= \frac{0.197 \times 0.500}{0.197 \times 0.500 + 0.102 \times 0.500} = 0.659$$

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

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

$$= \frac{0.898 \times 0.500}{0.898 \times 0.500 + 0.803 \times 0.500} = 0.528$$

FINAL COMMENTS:

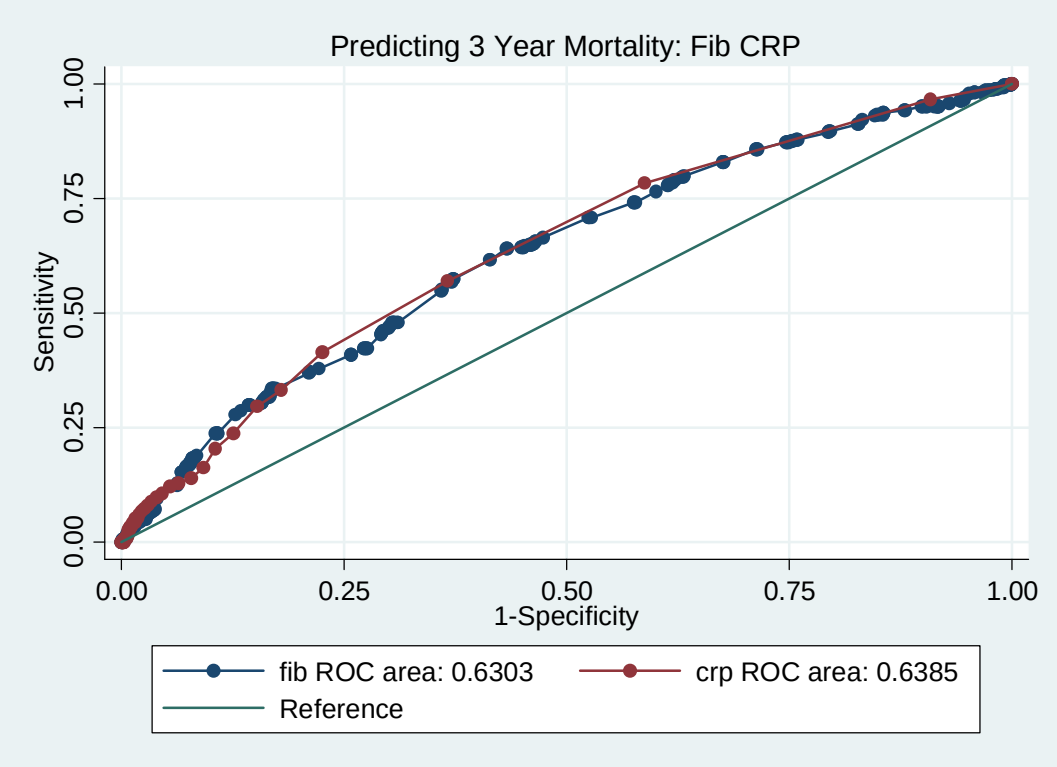
So we find that fibrinogen is a pretty bad test for predicting death. However, fibrinogen is highly associated with death. Using the 400mg/dl threshold and using death within 3 years, we find that 5.8% of the "low" fibrinogen group die within 3 years, and 13.9% of the "high" fibrinogen group die within 3 years. This 8.1% absolute difference in 3 year mortality is highly statistically significant (95% CI for absolute difference: 5.15% to 11.0%, two-sided P < 0.0001). This highlights the difference between detecting associations and predicting outcomes.

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. Below I have used the Stata command

roccomp deadIn3 fib crp, graph)

to produce ROC curves for both fibrinogen and CRP (see the key to Homework #5 from 2010 for similar analyses using CRP thresholds).

A completely worthless diagnostic procedure would have the sensitivity equal to 1 – specificity for all thresholds, and such a test would have an ROC curve on the 45 degree line. An ideal test would have 100% sensitivity while obtaining 100% specificity. Such a curve would be in the upper left hand portion of the ROC curve plot. A common summary measure for ROC curves is the "Area Under the Curve" (AUC), which is 0.5 for the worthless test and 1.0 for the ideal test. From the plot below, we find that both CRP and fibrinogen have AUCs around 0.63—not a very good value if you really needed to predict who was going to die.


Predicting 3 Year Mortality: Fib CRP
Sensitivity
0.00
0.25
0.50
0.75
1.00
1-Specificity
0.00
0.25
0.50
0.75
1.00
fib ROC area: 0.6303
crp ROC area: 0.6385
Reference