Biost 518: Applied Biostatistics II

Emerson, Winter 2006

Homework #4 Key

February 6, 2006

A separate file containing the Stata code and output used to solve this homework is available on the web pages.

Written problems due at the beginning of class, Monday, February 6, 2006.

The following questions pertain to the dataset for biomarkers of inflammation and cardiovascular disease stored as inflamm.txt on the class web page. For all questions involving statistical inference, provide estimates, confidence intervals, and P values in text suitable for a scientific journal.

1. We are interested in examining how mean C reactive protein levels vary by age and sex.
   a. Provide suitable descriptive statistics regarding the distribution of C reactive protein levels by age and sex.

Ans: From the following table and graph, we see a tendency for declining mean C reactive protein levels with increasing age in females, but a tendency for increasing mean C reactive protein levels with increasing age in males.

Table 1: Descriptive statistics for C reactive protein by sex and within age strata for each sex.

	N	Mean	SD	Min	25%ile	Median	75%ile	Max
Females	2861	3.62	5.48	0.07	0.98	2.02	3.59	86.36
65- 69 yo	1054	3.91	6.40	0.07	1.02	2.12	3.71	86.36
70- 74 yo	889	3.51	4.63	0.07	1.12	2.11	3.65	62.79
75- 79 yo	581	3.42	5.04	0.10	0.93	1.94	3.35	50.06
80- 84 yo	244	3.59	5.74	0.16	0.78	1.67	3.41	41.77
85- 89 yo	74	2.76	3.01	0.21	0.70	1.48	3.06	12.58
90-100 yo	19	1.86	1.28	0.31	0.86	1.48	3.08	4.55
Males	2072	3.58	6.96	0.07	0.92	1.76	3.11	107.97
65- 69 yo	644	3.47	7.55	0.19	0.93	1.76	3.07	107.97
70- 74 yo	688	3.14	5.10	0.07	0.93	1.72	2.98	47.82
75- 79 yo	431	3.90	7.33	0.20	0.89	1.78	3.49	82.95
80- 84 yo	219	4.09	7.55	0.26	0.87	1.76	3.09	51.45
85- 89 yo	70	4.54	8.12	0.17	0.91	1.98	3.88	53.14
90-100 yo	20	6.51	16.62	0.39	0.73	2.37	3.34	75.97

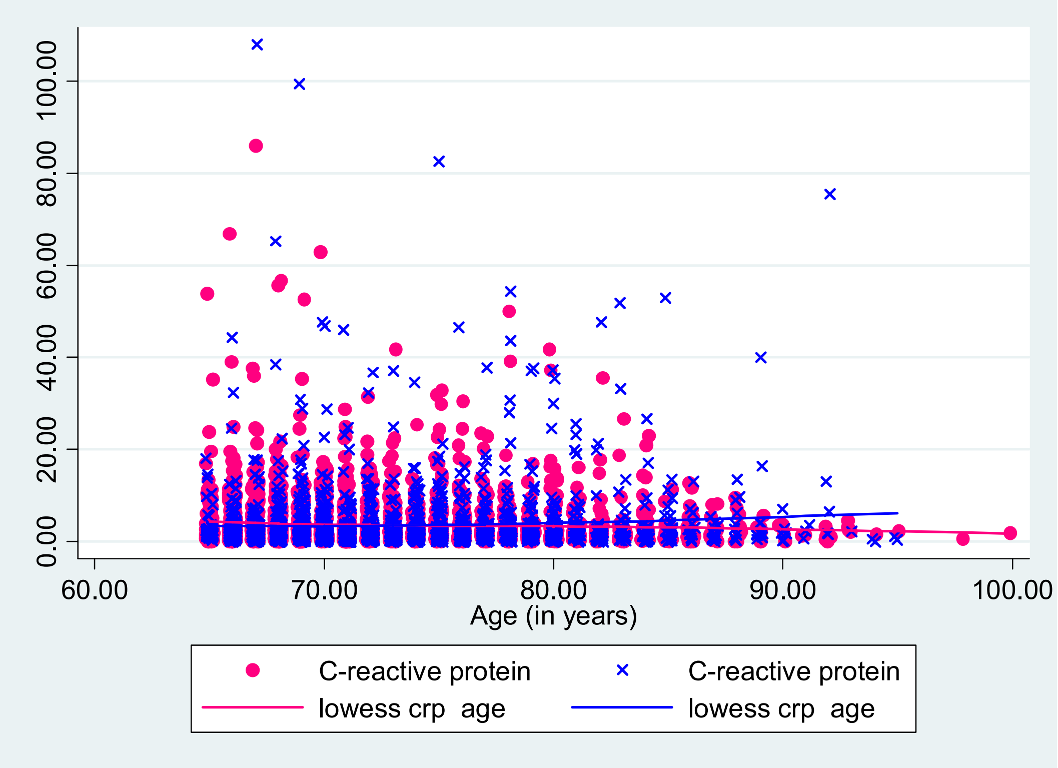

Figure 1: Scatterplot of C-reactive protein by age with superimposed lowess smooths for each sex (0= female, X= male).

b. Perform an analysis to determine whether the mean C reactive protein levels differ across sex groups.

<u>Ans</u>: On average, the mean C reactive protein levels in males is estimated to be 0.0399 mg/dl lower than in females (95% CI 0.401 mg/dl lower to 0.321 higher), a result that is not unexpected when there is no true difference between the sexes with respect to mean C reactive protein levels (P = 0.828).

c. Perform an analysis to determine whether there is a linear trend in mean C reactive protein levels by age.

<u>Ans</u>: On average, the mean C reactive protein levels is estimated to be 0.0007 mg/dl higher for every year difference in age, with the older subjects having higher levels (95% CI 0.035 mg/dl lower to 0.037 higher). This result is not unexpected when there is no true difference with respect to mean C reactive protein levels across age groups (P = 0.969).

d. Perform an analysis to determine whether the mean C reactive protein levels differ across sex groups after adjustment for age.

<u>Ans</u>: On average, the mean C reactive protein levels in males is estimated to be 0.040 mg/dl lower than in females of the same age (95% CI 0.397 mg/dl lower to 0.317 higher), a result that is not unexpected when there is no true difference between the sexes with respect to mean C reactive protein levels (P = 0.824).

e. Perform an analysis to determine whether there is a linear trend in mean C reactive protein levels by age after adjustment for sex.

<u>Ans</u>: On average, the mean C reactive protein levels is estimated to be 0.0009 mg/dl higher for every year difference in age when comparing subjects of the same sex (95% CI 0.035 mg/dl lower to 0.037 higher), with the older subjects tending toward higher levels. This result is not unexpected when there is no true difference with respect to mean C reactive protein levels across age groups (P = 0.960).

f. Perform an analysis to determine whether there is a linear trend in mean C reactive protein levels by age in women.

<u>Ans</u>: On average in women, the mean C reactive protein levels is estimated to be 0.051 mg/dl lower for every year difference in age (95% CI 0.087 mg/dl lower to 0.016 lower), with the older subjects tending to lower levels. This result is highly unusual when there is no true difference with respect to mean C reactive protein levels across age groups (P = 0.005).

g. Perform an analysis to determine whether there is a linear trend in mean C reactive protein levels by age in men.

<u>Ans</u>: On average in men, the mean C reactive protein levels is estimated to be 0.069 mg/dl higher for every year difference in age (95% CI 0.0013 mg/dl higher to 0.137 higher), with the older subjects tending to higher levels. This result is somewhat unusual when there is no true difference with respect to mean C reactive protein levels across age groups (P = 0.046).

h. Perform an analysis to test whether the results obtained in part g are statistically significantly different from those in part f. Interpret all parameters in the model used to answer this question, and relate those estimates to the parameter estimates obtained in parts f and g.

<u>Ans</u>: On average in women, the mean C reactive protein levels is estimated to be 0.051 mg/dl lower for every year difference in age (95% CI 0.087 mg/dl lower to 0.016 lower), with the older subjects tending to lower levels. On average in men, on the other hand, the mean C reactive protein levels is estimated to be 0.069 mg/dl higher for every year difference in age (95% CI 0.0013 mg/dl higher to 0.137 higher), with the older subjects tending to higher levels. This difference between the sexes in age effects of 0.120 (95% CI 0.0436 to 0.196) is beyond that which might be reasonably expected when the associatin between mean C reactive protein levels and age is the same for both sexes (P= 0.002).

(Note that I fit a regression model including an age-sex interaction in order to address this question. However, the estimates from that model are exactly the same as were estimated in the "disaggregated" analyses of the age effect in each sex separately. Furthermore, the statistical significance of the interaction is nearly identical to that we would have obtained has we used the regression results for the stratified analyses to create a statistic testing the difference in the slopes. That is, in the regressions used to answer parts f and g, the slope estimates were approximately normally distributed. Hence we could estimate the standard error of the difference between the slope estimates by taking the square root of the sum of the squared standard errors from the two

analyses. Fitting the regression model with the interaction let me avoid having to go to that trouble.)

i. How would you summarize the association between C reactive protein levels and age and sex?Provide a summary of your findings suitable for inclusion in a manuscript.

<u>Ans</u>: We found that there was, on average, no association between sex and mean C reactive protein levels when we averaged over the age distribution. Similarly, we found no association between age and mean C reactive protein levels when averaging across sexes. However, we did find that the effect of age on mean C reactive protein levels was modified by sex: Older women tended to have lower C reactive protein levels, and older men tended to have higher C reactive protein levels. This is, in my experience, a most unusual finding that significant trends would be in opposite directions in the two sexes. It is consistent with the hypothesis that, on average, C reactive protein levels tend to increase with age, but that such an increase in women is associated with an increased risk of death. Under this hypothesized scenario, the lower mean C reactive protein levels in older women might be due to "survivorship": Only women with lower C reactive protein levels tended to survive long enough to be sampled in this cross-sectional study. We can not, of course, prove that that mechanism is the true explanation for this observational data, and it does not agree with the general observation that men do not survive as long as women.

2. We are interested in examining how mean C reactive protein levels vary across groups defined by cholesterol level.
   a. Provide suitable descriptive statistics regarding the distribution of C reactive protein levels across groups defined by cholesterol levels.

<u>Ans</u>: From the following table and graph, we see a tendency for a U shaped function in mean C reactive protein levels across groups defined by cholesterol. Groups with the lowest cholesterol have the worst 8 year survival probabilities, and groups in the 220 – 240 mg/dl range appear to have the lowest average C reactive protein levels. There is, however much variability in the C reactive protein levels, with a tendency toward a highly positively skewed distribution at all levels of cholesterol.

Table 2: Descriptive statistics for C reactive protein within strata defined by cholesterol levels.

Cholesterol Level	N	Mean	SD	Min	25%ile	Median	75%ile	Max
< 160 mg/dl	404	5.32	11.20	0.10	0.75	1.90	3.64	99.71
160-180 mg/dl	586	4.31	8.46	0.07	0.97	1.94	3.53	107.97
180-200 mg/dl	960	3.44	5.33	0.10	0.89	1.81	3.235	53.96
200-220 mg/dl	1028	3.45	5.24	0.10	0.95	1.88	3.43	66.47
220-240 mg/dl	859	2.91	3.83	0.07	0.96	1.82	3.22	62.79
240-260 mg/dl	541	3.58	5.23	0.16	1.10	2.11	3.51	56.83
> 260 mg/dl	552	3.23	3.99	0.17	1.03	2.03	3.23	29.61

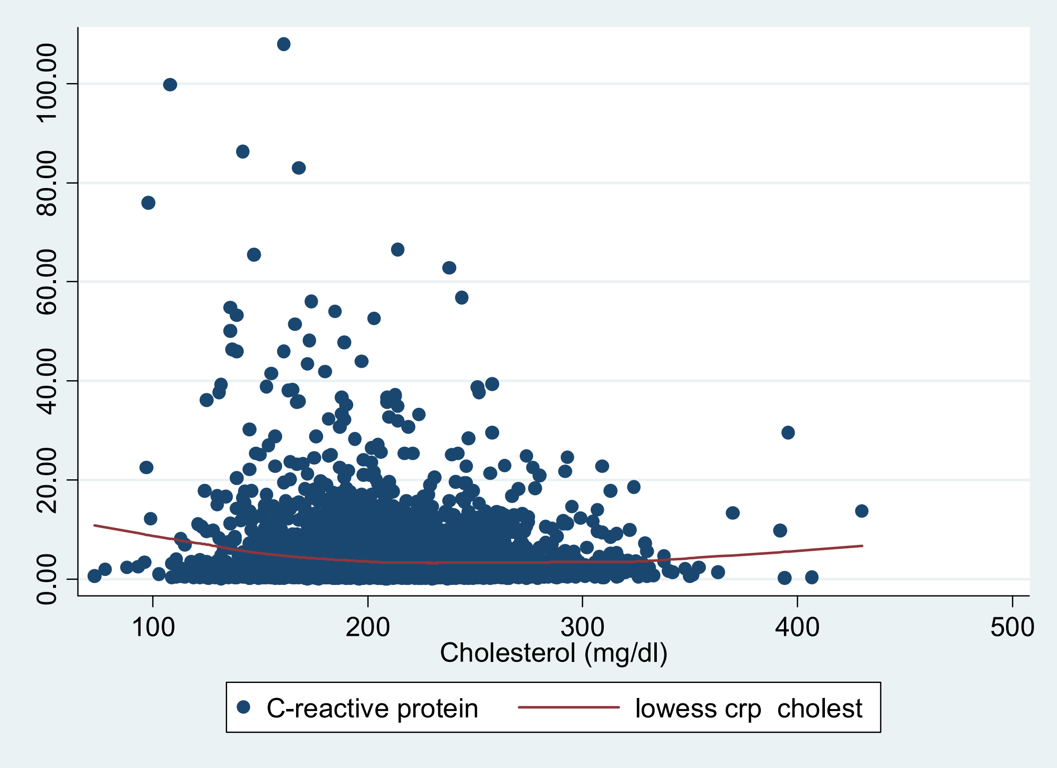

Figure 2: Scatterplot of C reactive protein versus cholesterol, with superimposed lowess smooth.

b. Perform an analysis to determine whether there is a linear trend in mean C reactive protein levels by cholesterol level.

Ans: On average, the mean C reactive protein levels is estimated to be 0.0121 mg/dl lower for every mg/dl difference in cholesterol (95% CI 0.006 mg/dl lower to 0.018 lower), with the subjects having higher cholesterol tending toward lower C reactive protein levels. This result is not unexpected when there is no true difference with respect to mean C reactive protein levels across age groups (P < 0.0005).

c. Perform an analysis to determine whether any trend in mean C reactive protein levels by cholesterol is well described by a straight line. That is, perform a test to see whether there is sufficient evidence in the data to suggest a nonlinear trend in mean C reactive protein levels by cholesterol. (A typical approach is to consider the possibility of a curvilinear trend by fitting both cholesterol and a new variable equal to the square of cholesterol.)

Ans: In linear regression modeling mean C reactive protein levels as a quadratic function of serum cholesterol, we find a statistically significant contribution from the squared cholesterol term (P < 0.0005). We thus conclude that we have evidence that the relationship between C reactive protein levels and cholesterol is nonlinear. From the quadratic model, we estimate an upward U-shaped function with a minimum occurring at approximately 245 mg/dl, however, the quadratic model may not be an accurate portrayal of the true relationship, either.

3. We are interested in examining how the distribution of time to death differs across groups defined by cholesterol level.
    a. Provide suitable descriptive statistics regarding the distribution of times to death across groups defined by cholesterol levels.

<u>Ans</u>: From the following tables and graphs, we see a tendency for a U shaped function in survival across groups defined by cholesterol. Groups with the lowest cholesterol have the worst 8 year survival probabilities, and groups in the 220 – 260 mg/dl range appear to have the best survival probabilities.

Table 3: Suvival probabilities at 1, 2, 5, and 8 years within strata defined by cholesterol level.

Cholesterol Stratum	N	Min Chol	Max Chol	Mean Chol	1yr Surv	2yr Surv	5yr Surv	8yr Surv
0: < 160 mg/dl	406	73	159	143.59	0.968	0.943	0.785	0.606
1: 160-180 mg/dl	591	160	179	170.64	0.975	0.949	0.834	0.689
2: 180-200 mg/dl	963	180	199	190.24	0.987	0.964	0.845	0.738
3: 200-220 mg/dl	1033	200	219	209.80	0.981	0.957	0.875	0.728
4: 220-240 mg/dl	862	220	239	229.09	0.987	0.970	0.904	0.805
5: 240-260 mg/dl	543	240	259	248.54	0.983	0.960	0.877	0.799
6: > 260 mg/dl	555	260	430	282.86	0.989	0.968	0.890	0.775

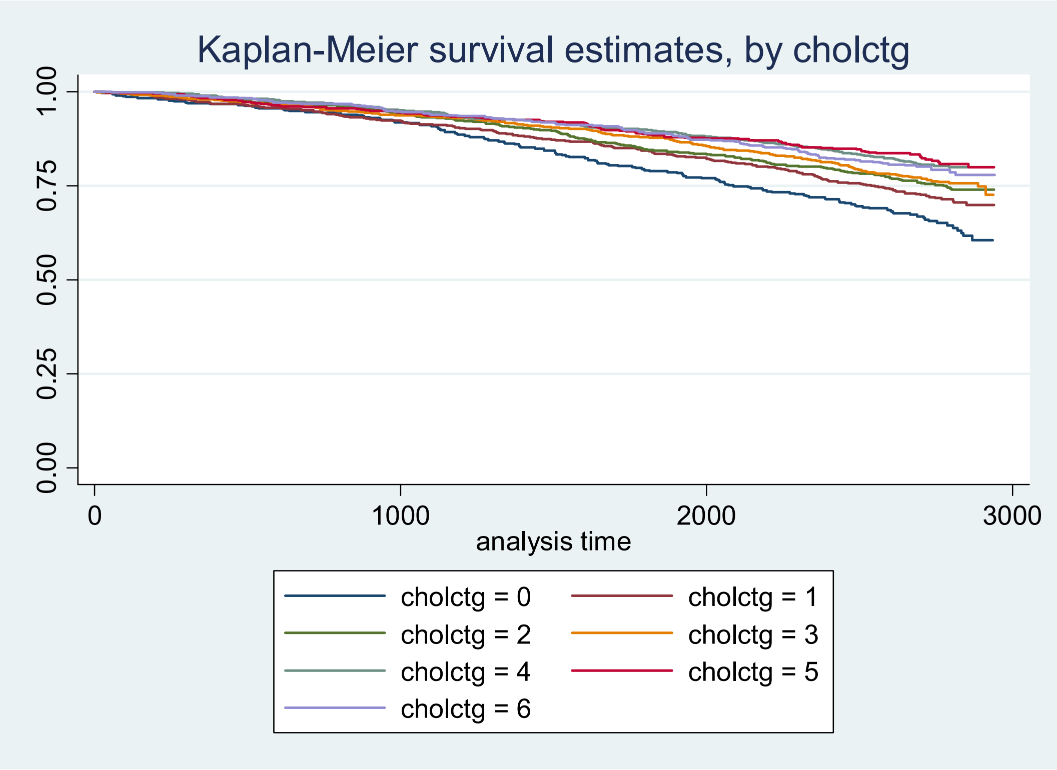

Figure 3: Kaplan-Meier survival curve estimates within strata defined by cholesterol level (see Table 3).

    b. Perform an analysis to determine whether there is a linear trend in the log hazard ratio by cholesterol level.

Ans: On average, the instantaneous risk of death is estimated to be 0.543% lower for every mg/dl difference in cholesterol (95% CI 0.387% lower to 0.699% lower), with the subjects having higher cholesterol tending toward lower risk of death. Thus for subjects differing by 50 mg/dl in their serum cholesterol, the group with the higher cholesterol is estimated to have a 23.837% lower risk of death (95% CI 17.642% to 29.570% lower). This result is highly unusual when there is no true difference with respect to risk of death across cholesterol groups (P < 0.0005).

c. Perform an analysis to determine whether any trend in the log hazard ratio by cholesterol is well described by a straight line. In particular, we are interested in seeing whether persons with low cholesterol as well as high cholesterol might be at increased risk of death compared to subjects with moderate cholesterol (say 200 mg/dl). (In working this problem, you may find it useful to create "centered" cholesterol variables defined relative to 200 mg/dl:
   - .g cenchol = cholest – 200
   - .g cenchol2 = cenchol^2

Ans: In proportional hazards regression modeling the log hazard function as a quadratic function of serum cholesterol, we find a statistically significant contribution from the squared cholesterol term (P = 0.001). We thus conclude that we have evidence that the relationship between instantaneous risk of death and cholesterol is nonlinear. From the quadratic model, we estimate an upward U-shaped function for the hazard ratio as shown in Figure 4 with a minimum occurring at approximately 284 mg/dl, however, the quadratic model may not be an accurate portrayal of the true relationship, either. From Table 3, we see that the best 2, 5, and 8 year survival probability tends to be in the 220 – 240 mg/dl group (and this is also subject to sampling variation).

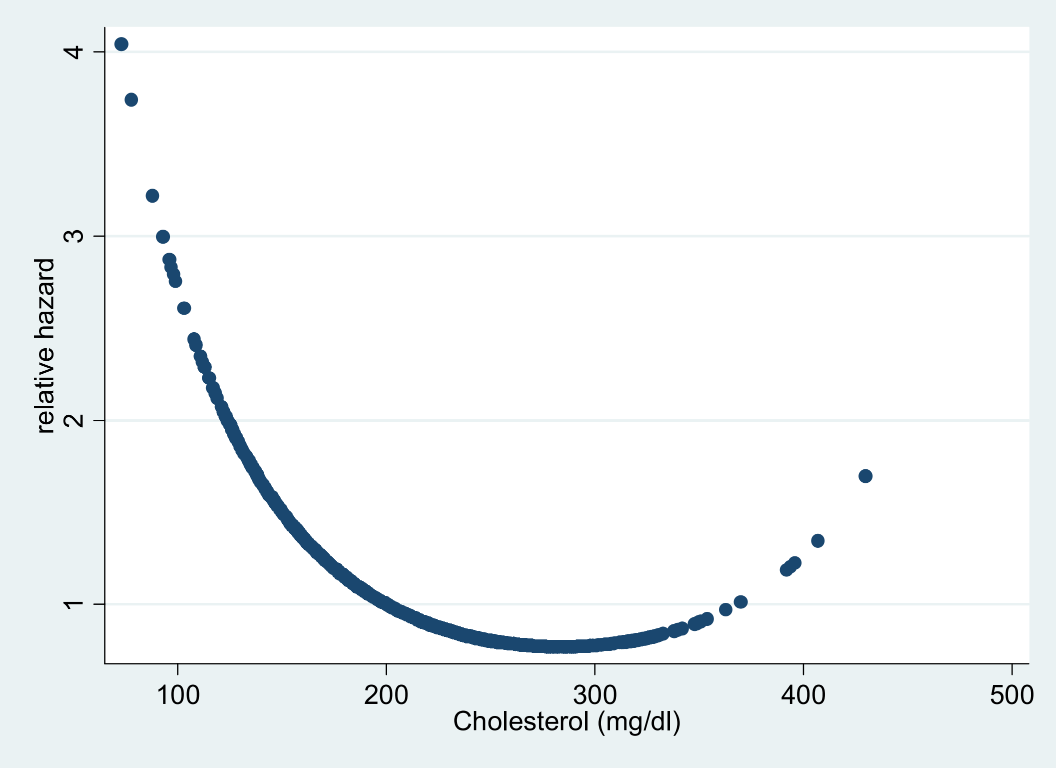

Figure 4: Estimated hazard ratio from a quadratic model of cholesterol as a function of cholesterol level. Hazard ratios are measured relative to a population having a cholesterol level of 200 mg/dl.

4. We are interested in examining how the distribution of time to death differs across groups defined by the inflammatory biomarkers after adjustment for age and sex.

   a. Provide suitable descriptive statistics regarding the distribution of times to death across groups defined by C reactive protein levels and fibrinogen levels.

Ans: From the following tables and graphs, we see a tendency for worse survival in groups having higher C reactive protein or fibrinogen levels.

Table 4: Survival probabilities at 1, 2, 5, and 8 years within strata defined by C reactive protein levels.

Stratum	N	Min C-RP	Max C-RP	Mean C-RP	1yr Surv	2yr Surv	5yr Surv	8yr Surv
0: < 0.5	428	0.07	0.49	0.37	1.00	0.98	0.91	0.79
1: 0.5-1	885	0.50	1.00	0.73	0.99	0.97	0.90	0.77
2: 1-2	1265	1.01	2.00	1.49	0.99	0.98	0.89	0.79
4: 2-4	1355	2.01	4.00	2.79	0.98	0.95	0.86	0.73
8: 4-8	486	4.01	7.99	5.82	0.96	0.92	0.79	0.64
16: > 8	514	8.01	107.97	16.48	0.97	0.92	0.79	0.62

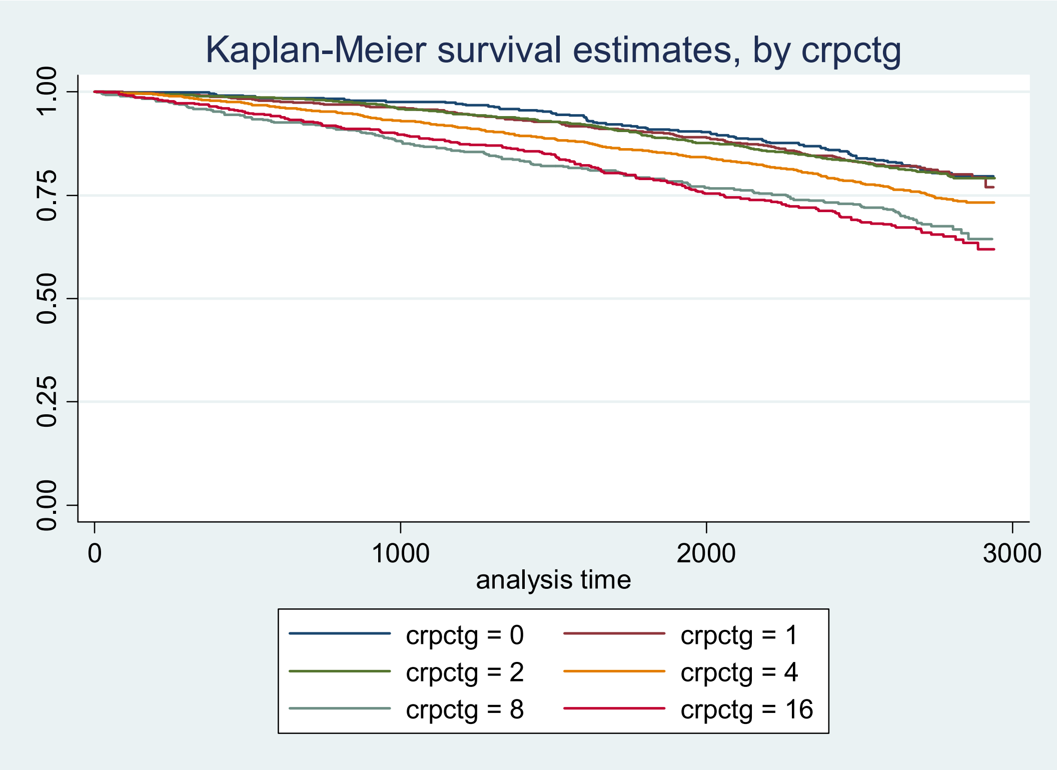

Figure 5: Kaplan-Meier estimates of survival within strata defined by C reactive protein levels (see Table 4).

Table 5: Survival probabilities at 1, 2, 5, and 8 years within strata defined by fibrinogen levels.

Fibrinogen Stratum	N	Min Fibr	Max Fibr	Mean Fibr	1yr Surv	2yr Surv	5yr Surv	8yr Surv
250	479	109	249	223.87	0.99	0.98	0.90	0.76
275	692	250	274	262.13	0.99	0.98	0.91	0.80
300	673	275	299	286.22	0.99	0.97	0.88	0.75
325	797	300	324	308.69	0.99	0.97	0.89	0.77
350	776	325	349	333.83	0.98	0.96	0.86	0.75
375	631	350	374	361.84	0.98	0.96	0.85	0.72
400	298	375	398	390.46	0.96	0.94	0.83	0.71
450	569	400	872	450.66	0.95	0.91	0.77	0.60

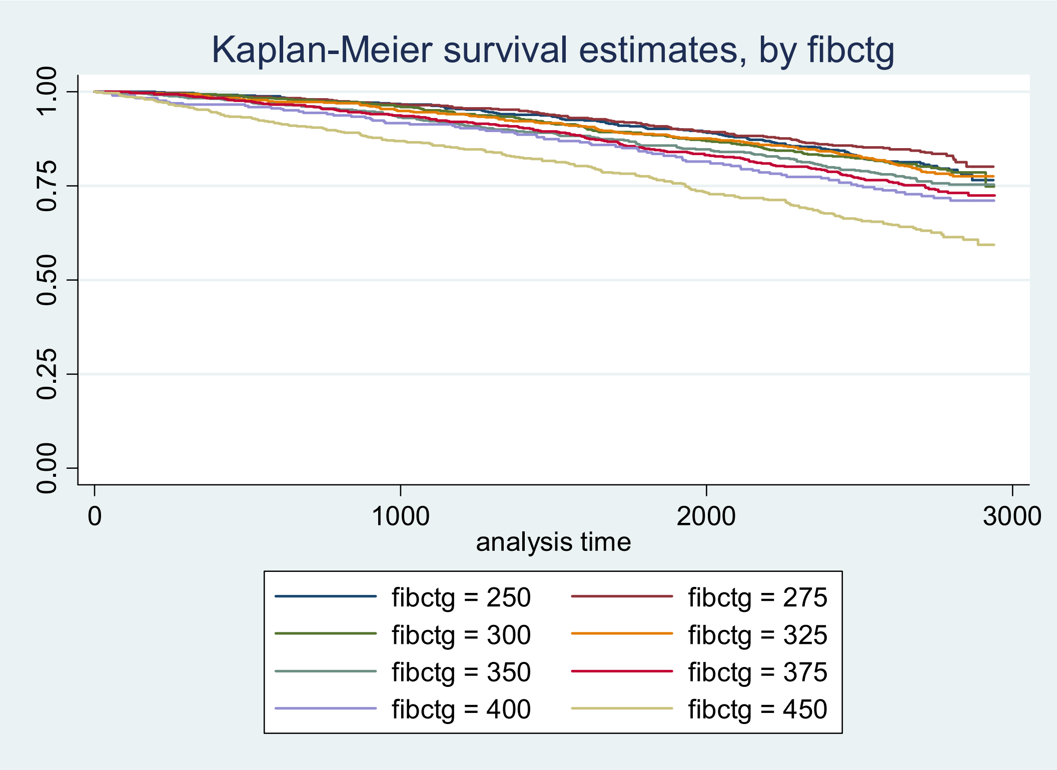

Figure 6: K aplan-Meier estimates of survival within strata defined by fibrinogen levels (see Table 5).

b. Perform an analysis to determine whether C reactive protein levels are associated with all cause mortality after adjustment for age and sex.

<u>Ans</u>: On average, the instantaneous risk of death is estimated to be 2.05% higher for every mg/l difference in C reactive protein when comparing groups of the same sex and age (95% CI 1.29% higher to 2.82% higher), with the subjects having higher C reactive protein levels tending toward higher risk of death. Thus for subjects of the same sex and age but differing by 5 mg/l in their C reactive protein level, the group with the higher C reactive protein level is estimated to have a 10.7% higher risk of death (95% CI 6.64% to 14.9% higher). This result is highly unusual when there is no true difference with respect to risk of death across cholesterol groups (P < 0.0005).

c. Perform an analysis to determine whether fibriniogen levels are associated with all cause mortality after adjustment for age and sex.

<u>Ans</u>: On average, the instantaneous risk of death is estimated to be 0.33% higher for every mg/dl difference in fibrinogen when comparing groups of the same sex and age (95% CI 0.25% higher to 0.41% higher), with the subjects having higher fibrinogen levels tending toward higher risk of death. Thus for subjects of the same sex and age but differing by 100 mg/dl in their fibrinogen level, the group with the higher fibrinogen level is estimated to have a 39.0% higher risk of death (95% CI 28.2% to 50.8% higher). This result is highly unusual when there is no true difference with respect to risk of death across cholesterol groups (P < 0.0005).

d. Perform an analysis to determine whether C reactive protein levels are associated with all cause mortality after adjustment for fibrinogen levels, age and sex. How would you explain any differences in your answers to parts b and d?

Ans: On average, the instantaneous risk of death is estimated to be 0.91% higher for every mg/l difference in C reactive protein when comparing groups of the same sex, age and fibrinogen levels (95% CI 0.03% higher to 1.79% higher), with the subjects having higher C reactive protein levels tending toward higher risk of death. Thus for subjects of the same sex, age, and fibrinogen levels but differing by 5 mg/l in their C reactive protein level, the group with the higher C reactive protein level is estimated to have a 4.61% higher risk of death (95% CI 0.15% to 9.27% higher). This result is somewhat unusual when there is no true difference with respect to risk of death across cholesterol groups (P = 0.042).

The marked change in the magnitude of the hazard ratio estimate for the C reactive protein level is due to confounding between fibrinogen and C reactive protein, supposing we did not consider fibrinogen to be in the causal pathway of interest. While we can not always rely on differences between the unadjusted and adjusted estimates to indicate confounding in logistic and proportional hazards regression, in this case we can: The adjusted estimate is closer to the null than the unadjusted estimate. Failure to adjust for a precision variable always attenuates the odds ratios or hazard ratios towards the null.

Had this not been the case, we would have had to rely on our definition of confounding. Certainly, fibrinogen levels are associated with risk of death as noted in part c. Then we can examine the correlation between fibrinogen levels and C reactive protein levels. The sample correlation of 0.48 suggests a reasonably strong association between these two variables. Thus, providing we did not consider fibrinogen in the causal pathway (a hypothesis which might, indeed, be entertained by some researchers), we would consider fibrinogen a confounder.