Biost 517
Applied Biostatistics I

Final Examination Key
December 14, 2005

Name: _____

Instructions: Please provide concise answers to all questions. Rambling answers touching on topics not directly relevant to the question will tend to count against you. Nearly telegraphic writing style is permissible..

The examination is closed book and closed notes. You may use calculators, but you may not use any special programs written for programmable calculators. Should you not have a calculator available, write down the equation that you would plug into a calculator.

If you come to a problem that you believe cannot be answered without making additional assumptions, clearly state the reasonable assumptions that you make, and proceed.

Please adhere to and sign the following pledge. Should you be unable to truthfully sign the pledge for any reason, turn in your paper unsigned and discuss the circumstances with the instructor.

PLEDGE:

On my honor, I have neither given nor received unauthorized aid on this examination:

Signed:_____

All problems use a subset of data from the Cardiovascular Health Study. This large government sponsored cohort study followed more than 5,000 adults aged 65 years and older living in one of four communities. The primary goals of the study were to observe the incidence of cardiovascular disease (especially heart attacks and heart failure) and cerebrovascular disease (especially strokes) in the elderly over an 11 year period.

The data used in this examination are the following variables measured on 5,000 individuals.

- *ptid*= patient identification number uniquely identifying each participant
- *site*= clinical site for each participant (coded 1, 2, 3, or 4)
- *age*= age (in years) of the participant at the start of the study
- *male*= indicator that the participant is male (0= female, 1= male)
- *smoker*= indicator that the participant was a smoker at the start of the study(0= no, 1= yes)
- *cholest*= the particpant's serum cholesterol level at the start of the study (mg/dl)
- *obstime*= time of observation (in years) until death or last follow-up for the participant
- *dead*= indicator that the participant was dead at the time recorded in *obstime*

1. (20 points) The following table presents descriptive statistics for the dataset.

variable	N	mean	sd	min	p25	p50	p75	max
ptid	5000	2501	1444	1	1251	2501	3751	5000
site	5000	2.5	1.1	1.0	1.0	2.0	4.0	4.0
age	5000	72.8	5.6	65.0	68.0	72.0	76.0	100.0
male	5000	0.419	0.493	0.000	0.000	0.000	1.000	1.000
smoker	4994	0.121	0.326	0.000	0.000	0.000	0.000	1.000
cholest	4953	211.7	39.3	73.0	186.0	210.0	236.0	430.0
obstime	5000	6.484	1.853	0.014	5.593	7.331	7.682	8.055
dead	5000	0.224	0.417	0.000	0.000	0.000	0.000	1.000

   a. For each of the variables given above, indicate the descriptive statistics that are not of scientific use to answer <u>any</u> scientific question. <u>Very</u> briefly explain why. (Most often, a single word would suffice.)

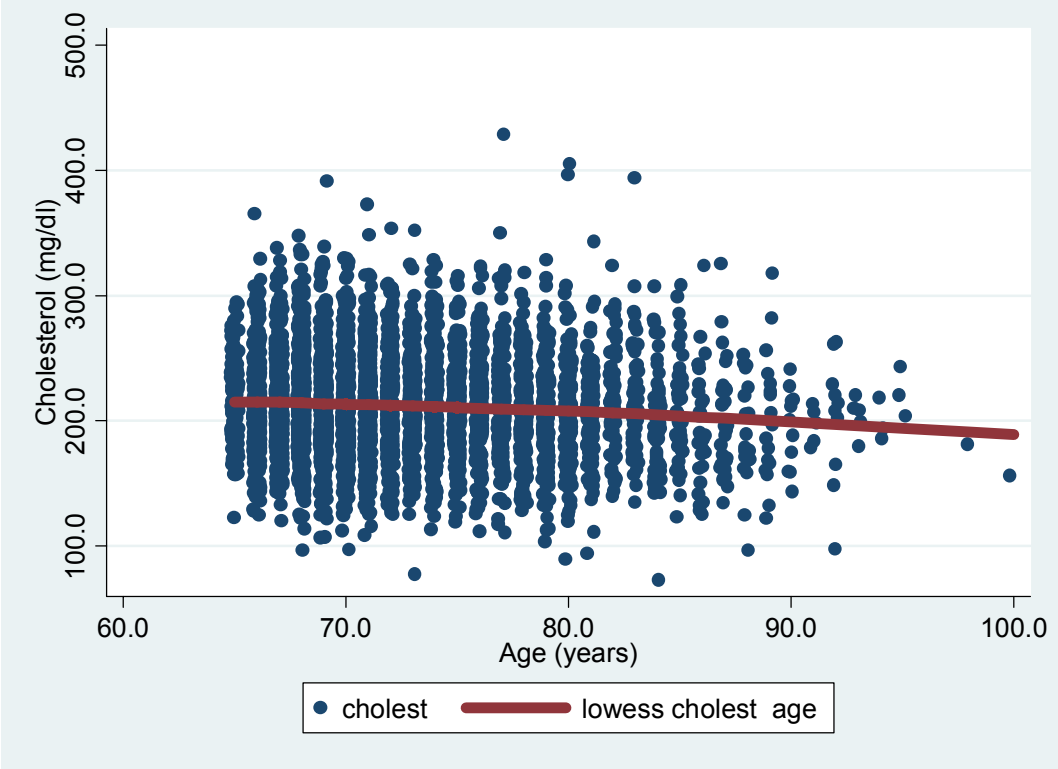
<u>Ans:</u> `ptid` and site are nominal (unordered categorical) variables, so none of the descriptive statistics are of interest. `male` and `smoker` are binary variables, so we can use the mean as the proportion of males and smokers in the sample. The other descriptive statistics are interpretable, but boring. `age` and `cholest` are continuous quantitative variables, so all of the descriptives are of interest. `obstime` and `dead` are variables measuring a censored time to event, so any scientifically useful descriptive statistics about the time to death or probability of surviving past any given time would need to be computed from Kaplan-Meier estimates.

   b. Using those descriptive statistics that are relevant, do any of the variables appear to be prone to outliers? <u>Very</u> briefly explain your reasoning.

<u>Ans:</u> Because the minimum age is only 1.5 SD below the median, while the maximum age is 5 SD above the median, I might worry that some of the oldest subjects are outliers. Similarly, for cholesterol, the minimum is 3.5 SD below the median, while the maximum is 5.5 SD above the median. This skewness may be due to outliers. (We cannot assess outliers very reliably in the censored variable, and the concept of outliers is not very interesting for binary variables or nominal variables.)

2. (36 points) The following analysis was performed to examine the relationship between age at study entry and serum cholesterol level at study entry.

Scatterplot of serum cholesterol by age (with superimposed lowess smooth)*:*



Linear regression analysis of serum cholesterol by age (with robust SE)*:*

```
. regress cholest age, robust
Linear regression                               Number of obs =    4953
                                                F(  1,  4951) =   30.73
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0063
                                                Root MSE      =  39.168
```

cholest	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	-.5589903	.1008394	-5.54	0.000	-.7566803	-.3613004
cons	252.3886	7.341587	34.38	0.000	237.9958	266.7813

a. Based on the above regression model, what is the best estimate for the mean cholesterol in 70 year old subjects?

Ans: 252.4 + 70 x (-0.5590) = 213.27 *(Note that the fact that the lowess smooth appears fairly linear makes me relatively comfortable in using the estimate from the linear regression model to predict group means for each age in the range sampled.)*

b. Based on the above regression model, what is the best estimate for the mean cholesterol in 71 year old subjects?

Ans: 252.4 + 71 x (-0.5590) = 212.711

c. Based on the above regression model, what is the best estimate for the mean cholesterol in 75 year old subjects?

Ans: 252.4 + 75 x (-0.5590) = 210.475

d. Based on the above regression model, what is the best estimate for the standard deviation of cholesterol measurements in a group of subjects who are all the same age?

Ans: 39.168 (obtained from Root MSE)

e. Based on the above regression model, what is the best estimate for the difference in mean cholesterol between 83 year old subjects and 82 year old subjects?

Ans: -0.559, with the older subjects having the lower mean (taken from the slope)

f. Based on the above regression model, what is the best estimate for the difference in mean cholesterol between 80 year old subjects and 90 year old subjects?

Ans: 10 x (-0.5590) = -5.590, with the older subjects having the lower mean

g. What does the above data analysis say about the change in cholesterol measurements as a person ages five years?

Ans: Nothing. This is a cross-sectional study. We cannot tell whether the trend toward lower cholesterol in older age groups is due to a trend toward decreasing cholesterol with aging in each subject, due to a "birth year cohort effect" in which subjects have fairly constant cholesterol since birth but there is a trend in which people born more recently have higher cholesterol, or due to "survivorship" in which subjects have fairly constant cholesterol over time, but those with higher cholesterol die early, leaving only subjects with lower cholesterol to be sampled when they are older. Or it could be some combination of the above.

h. Provide an interpretation for the intercept in the above regression model. What scientific use would you make of this estimate?

Ans: The model estimates that newborns would have average cholesterol of 252.4 mg/dl. This is extrapolating way outside our data, so I would make no use of this.

i. Provide an interpretation for the slope in the above regression model. What scientific use would you make of this estimate?

Ans: The model estimates that two groups differing in age by 1 year would (on average over the ages 65 -100) have a difference in mean cholesterol of -0.559 mg/dl, with the older group having the lower cholesterol.

j. Is there evidence that the slope is different from 0? State your evidence.

Ans: Yes. The p value testing for a zero slope is P < 0.0005.

k. Is there evidence of an association between cholesterol and age? Provide text suitable for inclusion in a scientific manuscript.

Ans: The model estimates that when two groups of patients that differ in age, the mean cholesterol is on average 0.559 mg/dl lower in the older group for each year difference in age. The 95% confidence interval suggests that these results are reasonably typical of what might be observed if the true association

between cholesterol and age were such that the mean cholesterol in an older group were anywhere between 0.361 mg/dl to 0.757 mg/dl per year difference in age. Thus we find a highly significant association between mean cholesterol and age (P < 0.0005).

l. The correlation between cholesterol and age was estimated to be r= -0.0793. Is that correlation statistically significantly different from 0? Briefly state your evidence.

Ans: Yes. The test for nonzero correlation is exactly the same as the test for nonzero slope, so we would use P < 0.0005 to conclude a highly statistically significant negative correlation between mean cholesterol and age.

3. (15 points) The following analyses were generated in order to estimate 4 year and 5 year survival probabilities for this cohort of patients.

Descriptive statistics for *obstime* according to whether death was observed:

```
. bysort dead: tabstat obstime, stat(n mean sd min p25 p50 p75 max) col(stat)

-> dead = 0.000
variable |    N   mean     sd    min    p25    p50    p75    max
 obstime | 3879  7.129  1.132  4.052  7.201  7.463  7.759  8.055

-> dead = 1.000
variable |    N   mean     sd    min    p25    p50    p75    max
 obstime | 1121  4.255  2.116  0.014  2.557  4.405  6.122  7.973
```

Creation of variables *obsgt4* and *obsgt5* dichotomizing *obstime* at 4 and 5 years:

```
. g obsgt4 = 0
. replace obsgt4= 1 if obstime > 4
. g obsgt5 = 0
. replace obsgt5= 1 if obstime > 5
```

Confidence intervals for the probability that *obstime* exceeds 4 or 5 years:

```
. ci obsgt4 obsgt5, binomial

                                                    -- Binomial Exact --
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
      obsgt4 |       5000       .9010    .0042237        .8923860    .9091425
      obsgt5 |       5000       .7676    .0059731        .7556386    .7792482
```

Kaplan-Meier estimates at 4 or 5 years:

```
. stset obstime dead
. sts list, at(4 5)
             Beg.                    Survivor      Std.
    Time     Total      Fail         Function      Error     [95% Conf. Int.]
-------------------------------------------------------------------------------
       4      4506       495           0.9010     0.0042     0.8924    0.9090
       5      3839       172           0.8634     0.0049     0.8534    0.8728
```

a. How do you explain the similarity between the analyses based on the dichotomized variable *obsgt4* and the Kaplan-Meier estimate of 4 year survival probability?

Ans: The earliest censored observation occurred at 4.052 years. With uncensored data, the Kaplan-Meier estimates agree exactly with the sample proportion, so

for any time prior to 4.052 years, estimated survival probabilities using dichotomized data will agree with the survival probability estimated using the Kaplan-Meier estimates.

b. How do you explain the differences between the analyses based on the dichotomized variable *obsgt5* and the Kaplan-Meier estimate of 5 year survival probability?

Ans: The earliest censored observation occurred at 4.052 years. Hence, some of the observation times that were less than 5 years were censored observations. In this setting, the dichotomized variable would not be informative about the true survival probability, while the Kaplan-Meier estimate would.

c. Would a two-sided level 0.05 hypothesis test reject a null hypothesis that the probability of surviving beyond 5 years is 86%? Very briefly justify your answer.

Ans: I would want to use an analysis based on the Kaplan-Meier estimates. While a P value was not explicitly given, I can use the fact that the 95% CI for five year survival probability include 0.86 to conclude that we would not reject that null hypothesis in a two-sided level 0.05 test. *(I could also compute a Z statistic using Z= (0.8634 – 0.86) / 0.0049 = 0.694, and as this is less than 1.96 in absolute value, we know we would not reject the null hypothesis.)*

4. (20 points) The following analyses explore the association between cholesterol and four year survival using the dichotomized observation time variable *obsgt4*.

Descriptive statistics for *cholest* and *obsgt4* by sex*:*

```
. bysort male: tabstat cholest obsgt4, stat(n mean sd min p25 p50 p75 max)
-> male = 0.000
variable |    N   mean     sd    min    p25    p50    p75    max
 cholest | 2870  221.5   38.9   88.0  195.0  219.0  245.0  430.0
  obsgt4 | 2904  0.933  0.251  0.000  1.000  1.000  1.000  1.000

-> male = 1.000
variable |    N   mean     sd    min    p25    p50    p75    max
 cholest | 2083  198.2   35.7   73.0  174.0  197.0  221.0  407.0
  obsgt4 | 2096  0.857  0.350  0.000  1.000  1.000  1.000  1.000
```

T test comparing cholesterol across groups defined by 4 year survival status*:*

```
. ttest cholest, by(obsgt4) unequal
Two-sample t test with unequal variances
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
       0 |     486     204.072    1.876692    41.37245    200.3846    207.7595
       1 |    4467     212.518    .5830693     38.9698    211.3749    213.6611
combined |    4953    211.6893    .5582482    39.28814    210.5949    212.7837
    diff |           -8.446005    1.965183               -12.30571   -4.586298
    diff = mean(0) - mean(1)                                      t =  -4.2978
Ho: diff = 0                     Satterthwaite's degrees of freedom =  582.562

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

T test comparing cholesterol across groups defined by 4 year survival status within for each sex separately*:*

```
. bysort male: ttest cholest, by(obsgt4) unequal

-> male = 0.000
```

```
Two-sample t test with unequal variances
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     193    217.3731    3.114907    43.27367    211.2292    223.5169
       1 |    2677    221.7916    .7442169    38.50558    220.3323    223.2509
---------+--------------------------------------------------------------------
combined |    2870    221.4944    .7252257    38.85207    220.0724    222.9164
---------+--------------------------------------------------------------------
    diff |           -4.418501    3.202577               -10.73105    1.894053
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -1.3797
Ho: diff = 0                     Satterthwaite's degrees of freedom =  214.495

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0846         Pr(|T| > |t|) = 0.1691          Pr(T > t) = 0.9154

-> male = 1.000
Two-sample t test with unequal variances
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     293    195.3106    2.199723    37.65319    190.9813    199.6399
       1 |    1790    198.6492    .8363772    35.38577    197.0088    200.2895
---------+--------------------------------------------------------------------
combined |    2083    198.1795    .7827123    35.72291    196.6446    199.7145
---------+--------------------------------------------------------------------
    diff |           -3.338582    2.353361               -7.965775    1.288611
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -1.4186
Ho: diff = 0                     Satterthwaite's degrees of freedom =  381.229

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
Pr(T < t) = 0.0784         Pr(|T| > |t|) = 0.1568          Pr(T > t) = 0.9216
```

a. Is there evidence of a statistically significant association between cholesterol and 4 year survival? Provide estimates and inferential statistics in support of your answer.

Ans: **Yes. From the t test comparing groups according to their vital status at 4 years, we find that subjects who survive beyond 4 years average cholesterol levels that are 8.45 mg/dl higher than subjects dying within 4 years of study entry (95% CI 4.59 to 12.3 mg/dl lower on average). Based on the two-sided P < 0.0001 from the t test which allows unequal variances, we conclude a highly statistically significant trend toward lower mean cholesterol among the survivors.**

b. Is there evidence that the association between cholesterol and 4 year survival is confounded by sex? Provide descriptive statistics in support of your answer.

Ans: **Yes. There appears to be an association between sex and survival beyond 4 years, because 93.3% of females survive 4 years, but only 85.7% of males survive for more than 4 years. (This fits in with what we know from other studies about the trend toward shorter longevity in males.) There also appears to be an association between sex and cholesterol levels in the sample, with men averaging cholesterol levels that are 23.3 mg/dl lower than for the women.** *(Because we are considering means, we can also note that the unadjusted estimate of 8.45 mg/dl higher mean cholesterol among the survivors is different than the estimates of 4.42 mg/dl higher and 3.34 mg/dl higher obtained in the females and males, respectively. I further note that even had we not been looking at means, the fact that the stratum specific estimates are less extreme than the unadjusted estimate could be taken as evidence of confounding. When considering odds ratios or hazard ratios, stratifying on a precision variable will produce in each stratum estimates that are more extreme than (and in the same direction as) the unadjusted*

estimate. Only confounding can make the unadjusted estimate be in the opposite direction of or more extreme than all stratum specific estimates.)

c. Is there evidence that the association between cholesterol and 4 year survival is modified by sex? Provide descriptive statistics in support of your answer.

Ans: The stratum specific estimates of differences in mean cholesterol are (to me) sufficiently similar that I would not regard that there was substantial effect modification by sex. *(If you considered 4.42 and 3.34 a meaningful difference, I gave you credit on this problem.)*

d. What statistic would you present to describe the association between cholesterol and 4 year survival? Provide the sentence you would use to report the results of your analysis.

Ans: Scientifically, I would regard that the presence of confounding by sex would mean that some adjustment should be made. (This is, of course, a judgment call.) I can compute a stratified estimate by taking the average of the estimates for the men and women (unweighted because approximately half the population is each sex): (4.42 + 3.34) / 2 = 3..88 mg/dl. Though I did not really expect you to do this calculation for the exam, the SE for this stratified estimate could be obtained as one-half the square root of the sum of the squared stratum specific SEs: sqrt(3.202577^2 + 2.353361^2) / 2 = 1.987. A 95% CI for the stratified estimate is thus 3.88 ± 1.96 x 1.987 = -0.0145 to 7.77 mg/dl. The Z statistic is 3.88 / 1.987 = 1.953. My complete report of this analysis would then be as follows (but I did not expect you to provide such a complete analysis in the exam):

> **The mean cholesterol is estimated to be 8.45 mg/dl higher among subjects who survive more than 4 years when compared to subjects dying within four years of study entry (95% CI 4.59 to 12.3 mg/dl lower). Although this difference was statistically significant (two-sided P < 0.0001), these results appeared to be confounded by sex: Males averaged cholesterol levels that were 23.3 mg/dl lower than females, and it has been well established that men have lower survival rates than women for myriad risk factors separate from cholesterol levels. When analyzed by each sex separately, females who survive more than 4 years had mean cholesterol 4.42 mg/dl higher than females who died within 4 years (95% CI 10.7 mg/dl lower to 1.89 mg/dl higher, two-sided P= 0.169), and males who survive more than 4 years had mean cholesterol 3.34 mg/dl higher than males who died within 4 years (95% CI 7.97 mg/dl lower to 1.29 mg/dl higher, two-sided P= 0.157). A stratified analysis of cholesterol level by vital status at 4 years adjusted for sex estimates that subjects who survive more than 4 years average cholesterol 3.88 mg/dl higher than subjects of the same sex dying within 4 years of study entry (95% CI 7.77 mg/dl lower to 0.0145 mg/dl higher). Because the 95% CI includes the null hypothesis of no difference, we were not able to establish an association between 4 year vital status and average cholesterol after adjusting for sex (P > 0.05).** *(While I could have done all of that analysis from the data provided, I could not have provided*

an exact P value without access to tables or a computer. The true P value from the above stratified analysis is P= 0.0508. For what it is worth, we do find from the Cardiovascular Health Study that subjects with higher cholesterol at baseline do appear to survive better than those with lower cholesterol. Many hypotheses have been put forward to explain why we might get different results in this older population than we get when we include adults in their 50s. My personal belief (which is not overly prejudiced by substantial knowledge on the subject) is that the better nutritional status associated with higher cholesterol confers protection against infectious diseases—a risk that increases with age, but is not so prominent among the 50 year olds. Among 50 year olds with high cholesterol, the deleterious effect of cholesterol on cardiovascular disease would outweigh any beneficial effect of fat reserves on infectious diseases.

5. (10 points) The following analyses explore the association between survival and cholesterol using proportional hazards regression.

```
. stset obstime dead
. stcox cholest, robust

Cox regression -- Breslow method for ties

No. of subjects      =         4953                 Number of obs    =      4953
No. of failures      =         1111
Time at risk         =  32191.17589
                                                    Wald chi2(1)     =     41.53
Log pseudolikelihood =   -9173.1039                 Prob > chi2      =    0.0000

------------------------------------------------------------------------------
             |               Robust
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     cholest |   .9945688   .0008405    -6.44   0.000     .9929228    .9962175
```

a. What conclusions would you reach about an association between cholesterol and survival based on this analysis. Provide a full description of your conclusions, including point estimate, confidence interval, and P value. Is any such association of a magnitude that would be scientifically important?

Ans: A proportional hazards analysis estimates that when comparing two groups which differ in their serum cholesterol levels, the risk of death is 0.543% lower for each 1 mg/dl difference in serum cholesterol, with better survival in the group with higher cholesterol (95% CI 0.378% lower to 0.708% lower for each 1 mg/dl difference in serum cholesterol). This observation is highly statistically significant (P < 0.0001). Such a difference is scientifically quite meaningful: According to these estimates, a 50 mg/dl difference in cholesterol levels (which corresponds to the difference between the 25th and 75th percentiles of the cholesterol levels in this sample) would be associated with a risk of death that was 23.8% lower in the group with higher cholesterol. *($0.9945688^{50} = 0.7616$, and $1 - 0.762 = 0.238 = 23.8\%$.)*

Grade distribution:

Total Possible: 101
Highest Achieved: 98
Mean: 81
SD: 13

Percentiles:

90th	80th	75th	70th	60th	50th	40th	30th	25th	20th
95	93	92.25	89.9	88	86	81	78	76	74.4