Biost 518
Applied Biostatistics II

Final Examination Key
March 15, 2006

Name: _____ Disc Sect: M W F

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

Problems 1-4 pertains to an analysis of 5,000 subjects participating in the Cardiovascular Health Study. Of particular interest is the association between obesity (as measured by body mass index (BMI: weight / height^2) and mortality. The results of various statistical analyses are given in the Appendices.

1. Appendix A contains descriptive statistics for the data. Consider the following univariate regression models exploring the association between survival and BMI. For each model give the interpretation of the slope parameter in terms of which summary measure is compared across groups. Indicate whether such an analysis would be valid in this dataset. (Note that you do not have results of these analyses. Instead just answer what you would be examining.)

    a. A linear regression of BMI (response) on a variable indicating a death observed within 4 years (predictor).

    <u>Ans</u>: The slope would estimate the difference in mean BMI between those subjects who die within 4 years and those subjects who survive more than 4 years after study entry. Though some subjects are censored in our data set, none are censored prior to 4 years, so we can do this analysis.

    b. A linear regression of a variable indicating a death observed within 4 years (response) on BMI (predictor).

    <u>Ans</u>: The slope would estimate the difference in the probability of surviving for more than 4 years between two groups that differ in BMI by 1 kg/m^2. Though some subjects are censored in our data set, none are censored prior to 4 years, so we can do this analysis. *(Note that this analysis is not the most popular way to compare distributions of binary variables across groups defined by a continuous predictor, but it would be valid so long as you use robust standard errors to account for the heteroscedasticity. While logistic regression is more common, the odds are less easily understood by naïve readers, so this analysis is in many ways preferable to my mind.)*

c. A logistic regression of a variable indicating a death observed within 4 years (response) on BMI (predictor).

Ans: The slope would estimate the difference in the log odds of surviving for more than 4 years between two groups that differ in BMI by 1 kg/m^2. This is of course equivalent to the log odds ratio comparing two such groups. Though some subjects are censored in our data set, none are censored prior to 4 years, so we can do this analysis.

d. A proportional hazards regression of obstime and death (response) on BMI (predictor).

Ans: The slope would estimate the log hazard ratio comparing two groups that differ in BMI by 1 kg/m^2. This is valid whether or not we have censored data, though it is nicer if we also have proportional hazards.

e. Very, very briefly, what are the relative advantages and disadvantages of these four analysis approaches?

Ans: The analyses based on the dichotomized death times would be expected to have less statistical power than the PH regression which uses the continuous measurement. There would not be too much difference between the statistical precision of the first three (parts a-c), so I would choose among them according to the ease of interpretation (difference in means or difference in probabilities wins here to my mind) and/or the correspondence to my presumed cause-effect relationship (using death as the response wins here).

For what it is worth, we can examine how those analyses would behave in this data set (variable death4 is as defined in Appendix B):

```
. regress bmi death4, robust
Linear regression                               Number of obs =    4987
                                                F(  1,  4985) =    2.95
                                                Prob > F      =  0.0862
                                                R-squared     =  0.0006
                                                Root MSE      =  4.7341
------------------------------------------------------------------------------
             |               Robust
         bmi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      death4 |  -.4023928   .2344803    -1.72   0.086    -.8620773    .0572917
       _cons |   26.70848    .070227   380.32   0.000      26.5708    26.84615
------------------------------------------------------------------------------

. regress death4 bmi, robust
Linear regression                               Number of obs =    4987
                                                F(  1,  4985) =    2.93
                                                Prob > F      =  0.0871
                                                R-squared     =  0.0006
                                                Root MSE      =  .29843
------------------------------------------------------------------------------
             |               Robust
      death4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |  -.0015991   .0009344    -1.71   0.087    -.003431    .0002328
        cons |   .1415022   .0255367     5.54   0.000     .091439    .1915654
------------------------------------------------------------------------------
```

```
. logistic death4 bmi, robust
Logistic regression                                     Number of obs   =       4987
                                                        Wald chi2(1)    =       2.75
                                                        Prob > chi2     =     0.0974
Log pseudolikelihood = -1606.9874                       Pseudo R2       =     0.0010
------------------------------------------------------------------------------
             |               Robust
      death4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |   .9816028   .0109956    -1.66   0.097     .9602866    1.003392
------------------------------------------------------------------------------

. stcox bmi, robust
Cox regression -- Breslow method for ties

No. of subjects      =         4987                     Number of obs   =      4987
No. of failures      =         1117
Time at risk         =  32340.49006
                                                        Wald chi2(1)    =      5.32
Log pseudolikelihood =   -9249.3938                     Prob > chi2     =    0.0211
------------------------------------------------------------------------------
             |               Robust
           t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |   .9835668   .0070677    -2.31   0.021     .9698115    .9975172
------------------------------------------------------------------------------
```

In linear regression, we find next to no difference in the P value whether bmi is response or the predictor. This makes sense given the correspondence between simple linear regression and correlation, and the fact that correlation treats the two variables symmetrically. The P value for logistic regression is not much different, but we do have more statistical significance with the PH regression. Of course, all of the slope estimates have different interpretations, so there is no comparability (except in sign) there.

2. Appendix B contains the results of a logistic regression performed on a variable indicating a death observed within 4 years on BMI, age, sex, and the BMI-sex and age-sex interactions. Use the results presented in Appendix B to answer the following questions.

   a. What is the estimated odds of death within 4 years for a 70 year old female with a BMI of 30 kg/m^2?

   Ans: -12.64 + 0.1245 × 70 + 0.0281 × 30 + 4.898 × 0 + (-0.03564) × 0 + (-0.05322) × 0 = -3.082 is the log odds, so exp(-3.082)= 0.04587.

   b. What is the estimated probability of death within 4 years for a 70 year old female with a BMI of 30 kg/m^2?

   Ans: 0.04587 / (1 + 0.04587) = 0.04386. *(Probability is odds / (1 + odds).)*

   c. What is the estimated odds of death within 4 years for a 71 year old female with a BMI of 30 kg/m^2?

   Ans: 0.04587 × 1.1326 = 0.05195. *(I just used the OR associated with a 1 year difference in age multiplied by my answer to part a. The other way to solve this would be: -12.64 + 0.1245 × 71 + 0.0281 × 30 + 4.898 × 0 + (-0.03564) × 0 + (-0.05322) × 0 = -2.9575 is the log odds, so exp(-2.9575)= 0.05195.)*

   d. What is the estimated odds of death within 4 years for a 70 year old female with a BMI of 31 kg/m^2?

Ans: 0.04587 × 1.0285 = 0.04718. *(I just used the OR associated with a 1kg/m^2 difference in BMI multiplied by my answer to part a. The other way to solve this would be: -12.64 + 0.1245 × 70 + 0.0281 × 31 + 4.898 × 0 + (-0.03564) × 0 + (-0.05322) × 0 = -2.9575 is the log odds, so exp(-2.9575)= 0.05195.)*

e. What is the estimated odds of death within 4 years for a 70 year old male with a BMI of 30 kg/m^2?

Ans: -12.64 + 0.1245 × 70 + 0.0281 × 30 + 4.898 × 1 + (-0.03564) × 70 + (-0.05322) × 30 = -2.275 is the log odds, so exp(-2.275)= 0.1028.

f. What is the estimated odds of death within 4 years for a 71 year old male with a BMI of 30 kg/m^2?

Ans: 0.1028 × 1.1326 × 0.9650 = 0.1124. *(I just used the OR associated with a 1 year difference in age multiplied by my answer to part e. Note that I had to consider both the main effect for age and the age-sex interaction. The other way to solve this would be: -12.64 + 0.1245 × 71 + 0.0281 × 30 + 4.898 × 1 + (-0.03564) × 71 + (-0.05322) × 30 = -2.1865 is the log odds, so exp(-2.1865)= 0.1123.)*

g. What is the estimated odds of death within 4 years for a 70 year old male with a BMI of 31 kg/m^2?

Ans: 0.1028 × 1.0285 × 0.9482 = 0.1003. *(I just used the OR associated with a 1 kg/m^2 difference in BMI multiplied by my answer to part e. Note that I had to consider both the main effect for BMI and theBMI-sex interaction. The other way to solve this would be: -12.64 + 0.1245 × 70 + 0.0281 × 31 + 4.898 × 1 + (-0.03564) × 70 + (-0.05322) × 31 = -2.3005 is the log odds, so exp(-2.3005)= 0.1002.)*

h. What is the interpretation of the intercept in the regression model computed using the logit command?

Ans: The log odds of death within four years for a newborn (age = 0) female having no weight but some height (so bmi = 0). *(Clearly scientifically absurd.)*

i. What is the interpretation of the slope parameter for age?

Ans: The log odds ratio comparing odds of death between groups of females having the same BMI but differing in age by 1 year.

j. What is the interpretation of the slope parameter for BMI?

Ans: The log odds ratio comparing odds of death between groups of females having the same age but differing in BMI by 1 kg/m^2.

k. What is the interpretation of the slope parameter for sex?

Ans: The log odds ratio comparing odds of death within four years for a newborn male having no weight but some height and a newborn female having no weight but some height. *(Clearly scientifically absurd.)*

l. What is the interpretation of the slope parameter for the age-sex interaction?

Ans: The difference between the log odds ratio comparing odds of death between groups of males having the same BMI but differing in age by 1 year and the log

odds ratio comparing odds of death between groups of females having the same BMI but differing in age by 1 year.

m. What is the interpretation of the slope parameter for the bmi-sex interaction?

Ans: The difference between the log odds ratio comparing odds of death between groups of males having the same age but differing in BMI by 1 kg/m^2 and the log odds ratio comparing odds of death between groups of females having the same age but differing in BMI by 1 kg/m^2.

n. Suppose we fit a logistic regression model of death4 on bmi and age using the logit command for just females. What would be the values of the estimated intercept, age slope parameter, and bmi slope parameter?

Ans: These parameters would correspond exactly to the estimated intercept, age slope parameter, and bmi slope parameter reported in Appendix B: -12.64, 0.1245, and 0.0281, respectively.

o. Suppose we fit a logistic regression model of death4 on bmi and age using the logit command for just males. What would be the values of the estimated intercept, age slope parameter, and bmi slope parameter?

Ans: These parameters would correspond to the respective sums of the estimated intercept plus the sex effect, the age slope parameter plus the age-sex interaction, and the bmi slope parameter plus the bmi-sex interaction reported in Appendix B: -12.64 + 4.898 = -7.742, 0.1245 – 0.03564 = 0.08886, and 0.0281 – 0.05322 = -0.0251, respectively.

3. Appendix C contains the results of a logistic regression analysis performed on a variable indicating a death observed within 4 years on BMI, age, sex, and the BMI-sex and age-sex interactions and using robust standard errors. Note that I fit that same model under two different parameterizations: Once based on an indicator of male sex and once on an indicator of female sex. Also included are various tests of combinations of regression parameters. Use the results presented in Appendix C to answer the following questions. **For each question make certain that you identify which test and p value you use to answer the question.**

   a. Is there evidence that there is a statistically significant difference between your answer to parts a and c in problem 2? (That is, is there evidence of an age effect in females of the same BMI?)

   Ans: Using the test of the slope parameter for age in the model including an indicator of male sex (so the age slope parameter corresponds to females), we find a highly statistically significant difference: P < 0.0005.

   b. Is there evidence that there is a statistically significant difference between your answer to parts a and d in problem 2? (That is, is there evidence of a BMI effect in females of the same age?)

   Ans: Using the test of the slope parameter for bmi in the model including an indicator of male sex (so the BMI slope parameter corresponds to females), we do not find a statistically significant difference: P = 0.073.

c. Is there evidence that there is a statistically significant difference between your answer to parts a and e in problem 2? (That is, is there evidence of a sex effect in subjects of similar age and BMI?)

Ans: The fact that the model includes interactions between sex and both age and BMI means that the question of a sex effect in subjects of similar age and BMI cannot be answered in general: It would have to be answered for each age-BMI combination separately. Among newborns having BMI=0, we can use the sex main effect to ascribe a statistically significant difference: P = 0.001. Appendix C does not contain enough information to be able to answer the specific question about 70 year olds having BMI of 30 kg/m^2.

(This was decidedly the hardest question on the exam. For your future reference I note that there are two ways we could get Stata to answer this question for us. The "test" command can be used after any regression command to test combinations of parameters. The difference between the answer to part a and part e is the sum of the male slope parameter, 70 times the male-age interaction, and 30 times the male-bmi interation. So after fitting the model shown in Appendix C we want to test whether that difference is 0:

```
. test male + 70*male_age + 30*male_bmi = 0
 ( 1)  male + 70 male_age + 30 male_bmi = 0
           chi2(  1) =   27.01
         Prob > chi2 =    0.0000
```

From the above, we find a highly statistically significant difference.

Another way would be to create new variables centering age at 70 and centering bmi at 30:

```
. g cage= age - 70
. g cbmi = bmi - 30
. g male_cage = male * cage
. g male_cbmi = male * cbmi
. logit death4 cage cbmi male male_cage male_cbmi, robust
Logistic regression                             Number of obs   =       4987
                                                Wald chi2(5)    =     243.21
                                                Prob > chi2     =     0.0000
Log pseudolikelihood = -1477.1073               Pseudo R2       =     0.0818
------------------------------------------------------------------------------
             |               Robust
      death4 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cage |   .1245051   .0118493    10.51   0.000     .1012809    .1477292
        cbmi |   .0280855   .0156766     1.79   0.073    -.0026401     .058811
        male |   .8067821   .1552499     5.20   0.000     .5024979    1.111066
   male_cage |  -.0356414    .015923    -2.24   0.025    -.0668499    -.004433
   male_cbmi |   -.053217   .0245816    -2.16   0.030    -.1013961   -.0050379
       _cons |  -3.078662   .1116874   -27.56   0.000    -3.297565   -2.859758
------------------------------------------------------------------------------
```

In the above model, the male slope parameter now compares males who are 70 years old with BMI of 30 kg/m^2 to females who are 70 years old with BMI of 30 kg/m^2 . We find that the log odds ratio is highly statistically significantly different from 0 (P < 0.0005) so the odds ratio is highly statistically different from 1.0.

Note that the square of 5.20 (the Z statistic for the male slope parameter in the model using the centered age and bmi variables) is 27.04 (the chi-squared statistic from the test of the linear combination of the slope parameters from modeling the uncentered variables as shown in Appendix C), so the two methods are really performing the exact

same test (there is only some round-off error causing the two statistics to disagree slightly).

d. Is there evidence that there is a statistically significant difference between your answer to parts e and f in problem 2? (That is, is there evidence of an age effect in males of the same BMI?)

Ans: Using the test of the slope parameter for age in the model including an indicator of female sex (so the age slope parameter corresponds to males), we find a highly statistically significant difference: P < 0.0005.

e. Is there evidence that there is a statistically significant difference between your answer to parts e and g in problem 2? (That is, is there evidence of a BMI effect in males of the same age?)

Ans: Using the test of the slope parameter for bmi in the model including an indicator of male sex (so the BMI slope parameter corresponds to males), we do not find a statistically significant difference: P = 0.184.

f. Is there evidence that the comparison between the odds estimated in parts a and d in problem 2 differs significantly from a comparison between the odds estimated in parts e and g in problem 2?

Ans: Using the test of the slope parameter for the male-bmi interaction in the model including an indicator of male sex, we do find a statistically significant difference: P = 0.030.

g. Is there evidence of an association between BMI and mortality within 4 years?

Ans: Using the test of the slope parameters for both the bmi slope parameter and the male-bmi interaction in the model including an indicator of male sex, we do not find a statistically significant difference: P = 0.0833.

h. Is there evidence of an association between sex and mortality within 4 years?

Ans: Using the test of the slope parameters for all three of the male slope parameter, the male-age interaction, and the male-bmi interaction in the model including an indicator of male sex, we do find a highly statistically significant difference: P < 0.00005.

4. The standard errors estimated in Appendix B (using classical logistic regression) and Appendix C (using robust standard errors) differ the most on a relative scale for BMI and the BMI-sex interaction.

a. What could best explain the reason for obtaining such different results?

Ans: Nonlinearity of the log odds of death within 4 years across groups defined by BMI. *(Classical logistic regression presumes that the variability of observations within groups will follow the mean-variance relationship of binary data when using the fitted values from the logistic regression model. Hence, unlike classical linear regression, classical logistic regression allows for heteroscedasticity when there is a difference in odds of events across groups. But classical logistic regression uses model based estimates of that heteroscedasticity. The robust standard errors will instead use an estimated variability similar to the sample variance. The only way these could differ*

markedly is if the model estimates odds (or probability) of death that differs from the observed data in a major way.)

b. Appendix D contains an analysis including quadratic terms for BMI and the BMI-sex interaction. Is there statistical evidence of a nonlinear association between BMI and the log odds of death within 4 years? Make clear your evidence for such an association.

Ans: Using the test of the quadratic term for BMI and the male interaction with the quadratic term for bmi, we find statistically significant evidence of departures from linearity: P = 0.0166. *(We would have to regard that a nonlinear association exists if it exists for either sex, so we test both slope parameters which involve a nonlinear function of BMI.)*

c. Is there evidence of an association between BMI and 4 year mortality using the results in Appendix D? Make clear what tests you use.

Ans: Using the test of all four parameters modeling the linear and quadratic association of BMI in females and the interaction between sex and those linear and quadratic terms, we find highly statistically significant evidence of an association between BMI and the odds of death within 4 years: P = 0.0016. *(We would have to regard that an association exists if it exists for either sex in either a linear or quadratic fashion, so we test all slope parameters which involve BMI in any way.)*

5. Suppose we are interested in exploring associations between systolic blood pressure and age and sex. Consider two possible study designs:

- Study A: Gather a single blood pressure measurement on 5,000 independent subjects of both sexes and ages between 60 and 80.

- Study B: Make five measurements one year apart (so a four year time span) on each of 1,000 independent subjects of both sexes and ages between 60 and 75.

a. Presuming appropriate statistical inference is performed in either case, which of Study A or B is more likely to provide more statistical precision to assess an association between blood pressure and age? Briefly explain why.

Ans: Study B. The same number of observations would be used in either case, so the answer depends upon how the correlated observations contribute to the comparison of groups. Blood pressure measurements are likely positively correlated within an individual. The longitudinal study will make comparisons across age groups in such a way that the repeated measurements made on the same individual will be in different age groups. Hence, we will in some sense be using differences of positively correlated observations, which tends to lead to smaller standard errors than differences between independent observations. *(The degree of improvement in the precision will depend upon how highly correlated the observations are within an individual, as well as the variability of ages that would result from the two studies.)*

b. Presuming appropriate statistical inference is performed in either case, which of Study A or B are more likely to provide more statistical precision to assess an association between blood pressure and sex? Briefly explain why.

Ans: Study A. Again, the same number of measurements are used in either study. Blood pressure measurements are likely positively correlated within an individual. The longitudinal study will make comparisons across sex groups in such a way that the repeated measurements made on the same individual will be in the same sex group. Hence, we will in some sense be using sums of positively correlated observations, which tends to lead to larger standard errors than sums of the same number of independent observations. *(The degree of loss of precision will depend upon how highly correlated the observations are within an individual.)*

APPENDIX A: Descriptive Statistics for problems 1-4

Problems 1-4 pertain to analyses of data 5,000 subjects participating in the Cardiovascular Health Study. Of particular interest is the association between mortality and body mass index (BMI: a measure of obesity). Data is available on the following variables :

- ***age*** = age in years
- ***male*** = indicator of male sex (0= female, 1= male)
- ***bmi*** = body mass index: weight / height^2 (kg/m^2)
- ***obstime*** = time in years between start of study and the earlier of the subject's death or the time of data analysis
- ***death*** = an indicator that the subject died at the time recorded in *obstime* (0= subject still alive at time of data analysis, 1= subject observed to die)

Descriptive statistics on the entire dataset:

. tabstat age male bmi obstime death, stat(n mean sd min q max) col(stat) format

variable	N	mean	sd	min	p25	p50	p75	max
age	5000	72.83	5.60	65.00	68.00	72.00	76.00	100.00
male	5000	0.419	0.493	0.000	0.000	0.000	1.000	1.000
bmi	4987	26.67	4.74	14.70	23.50	26.10	29.20	58.80
obstime	5000	6.484	1.853	0.014	5.593	7.331	7.682	8.055
death	5000	0.224	0.417	0.000	0.000	0.000	0.000	1.000

Descriptive statistics by sex:

Females:

. tabstat age bmi obstime death if male==0, stat(n mean sd min q max) col(stat) format

variable	N	mean	sd	min	p25	p50	p75	max
age	2904	72.56	5.52	65.00	68.00	71.00	76.00	100.00
bmi	2895	26.86	5.31	14.70	23.20	26.10	29.60	58.80
obstime	2904	6.648	1.682	0.063	6.664	7.348	7.682	8.055
death	2904	0.170	0.376	0.000	0.000	0.000	0.000	1.000

Males:

. tabstat age bmi obstime death if male==1, stat(n mean sd min q max) col(stat) format

variable	N	mean	sd	min	p25	p50	p75	max
age	2096	73.21	5.69	65.00	69.00	72.00	77.00	95.00
bmi	2092	26.41	3.78	15.60	23.90	26.10	28.50	46.20
obstime	2096	6.257	2.046	0.014	4.552	7.294	7.682	8.052
death	2096	0.299	0.458	0.000	0.000	0.000	1.000	1.000

Descriptive statistics of observation time by observed death:

. tabstat obstime, stat(n mean sd min q max) col(stat) format by(death)

obstime	N	mean	sd	min	p25	p50	p75	max
death= 0	3879	7.129	1.132	4.052	7.201	7.463	7.759	8.055
death= 1	1121	4.255	2.116	0.014	2.557	4.405	6.122	7.973

APPENDIX B: Logistic regression analysis of 4 year mortality rates

I generated a variable indicating a death observed within 4 years using Stata code

```
generate death4 = 0
replace death4 = 1 if obstime <=4 & death==1
```

I also generated variables modeling multiplicative interactions between BMI and sex and age and sex using Stata code

```
generate male_age = male * age
generate male_bmi = male * bmi
```

Logistic regression analyses were performed in Stata using both the "logit" command (so parameter estimates on the log odds scale) and using the "logistic" command (so odds ratio scale):

```
. logit death4 age bmi male male_age male_bmi
Logistic regression                               Number of obs   =       4987
                                                  LR chi2(5)      =     263.04
                                                  Prob > chi2     =     0.0000
Log likelihood = -1477.1073                       Pseudo R2       =     0.0818

------------------------------------------------------------------------------
      death4 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .1245051   .0119695    10.40   0.000     .1010453    .1479648
         bmi |   .0280855   .0140769     2.00   0.046     .0004953    .0556757
        male |   4.898193   1.436789     3.41   0.001     2.082138    7.714249
    male_age |  -.0356414   .0158928    -2.24   0.025    -.0667907   -.0044922
    male_bmi |   -.053217   .0226967    -2.34   0.019    -.0977018   -.0087323
        cons |  -12.63658   1.046067   -12.08   0.000    -14.68683   -10.58632
------------------------------------------------------------------------------
```

```
. logistic death4 age bmi male male_age male_bmi
Logistic regression                               Number of obs   =       4987
                                                  LR chi2(5)      =     263.04
                                                  Prob > chi2     =     0.0000
Log likelihood = -1477.1073                       Pseudo R2       =     0.0818

------------------------------------------------------------------------------
      death4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.132588   .0135565    10.40   0.000     1.106327    1.159472
         bmi |   1.028484   .0144778     2.00   0.046     1.000495    1.057255
        male |   134.0474   192.5979     3.41   0.001     8.021598     2240.04
    male_age |   .9649862   .0153363    -2.24   0.025      .935391    .9955179
    male_bmi |   .9481742   .0215204    -2.34   0.019     .9069193    .9913057
------------------------------------------------------------------------------
```

APPENDIX C: Logistic regression analysis of 4 year mortality rates

Using the same variables defined for Appendix B, I fit a logistic regression using robust standard error estimates. I provide test based on various combinations of the parameters. I then also fit a model in which I parameterized sex using female = 1 – male.

```
. logistic death4 age bmi male male_age male_bmi, robust
Logistic regression                               Number of obs   =       4987
                                                  Wald chi2(5)    =     243.21
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood = -1477.1073                 Pseudo R2       =     0.0818
------------------------------------------------------------------------------
             |               Robust
      death4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.132588   .0134203    10.51   0.000     1.106587    1.159199
         bmi |   1.028484   .0161231     1.79   0.073     .9973634    1.060575
        male |   134.0474   194.2494     3.38   0.001     7.830216     2294.79
    male_age |   .9649862   .0153654    -2.24   0.025     .9353356    .9955768
    male_bmi |   .9481742   .0233077    -2.16   0.030      .903575    .9949747
------------------------------------------------------------------------------

. test male_age male_bmi
 ( 1)  male_age = 0
 ( 2)  male_bmi = 0
           chi2(  2) =    8.51
         Prob > chi2 =    0.0142

. test  male male_age male_bmi
 ( 1)  male = 0
 ( 2)  male_age = 0
 ( 3)  male_bmi = 0
           chi2(  3) =   71.85
         Prob > chi2 =    0.0000

. test age male_age
 ( 1)  age = 0
 ( 2)  male_age = 0
           chi2(  2) =  180.20
         Prob > chi2 =    0.0000

. test bmi male_bmi
 ( 1)  bmi = 0
 ( 2)  male_bmi = 0
           chi2(  2) =    4.97
         Prob > chi2 =    0.0833

. logistic death4 age bmi female female_age female_bmi, robust
Logistic regression                               Number of obs   =       4987
                                                  Wald chi2(5)    =     243.21
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood = -1477.1073                 Pseudo R2       =     0.0818
------------------------------------------------------------------------------
             |               Robust
      death4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.092932    .011625     8.35   0.000     1.070383    1.115955
         bmi |   .9751816   .0184642    -1.33   0.184     .9396557    1.012051
      female |     .00746   .0108104    -3.38   0.001     .0004358    .1277104
  female_age |   1.036284   .0165007     2.24   0.025     1.004443    1.069135
  female_bmi |   1.054659   .0259252     2.16   0.030     1.005051    1.106715
------------------------------------------------------------------------------
```

APPENDIX D: Logistic regression analysis of 4 year mortality rates

Using the same variables defined for Appendix B, I generated variables modeling a quadratic effect of BMI and the interaction of that quadratic effect with sex. I fit a logistic regression using robust standard error estimates. I provide test based on various combinations of the parameters.

```
. g bmisqr= bmi^2
. g male_bmisqr= male * bmisqr

. logistic death4 bmi gender mbmi age mage bmisqr male_bmisqr, robust

Logistic regression                               Number of obs   =       4987
                                                  Wald chi2(7)    =     252.19
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood = -1473.3374                 Pseudo R2       =     0.0841

------------------------------------------------------------------------------
             |               Robust
      death4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |   .8512864   .0695751    -1.97   0.049      .725283    .9991802
        male |   128.1347   333.2151     1.87   0.062     .7836306    20951.83
         age |   1.132445   .0133928    10.52   0.000     1.106498    1.159001
    male_age |   .9646547   .0153618    -2.26   0.024     .9350111    .9952381
      bmisqr |   1.003145   .0012998     2.42   0.015       1.0006    1.005696
    male_bmi |   .9389369      .1466    -0.40   0.687     .6914088    1.275081
 male_bmisqr |   1.000523   .0027319     0.19   0.848     .9951828    1.005892
------------------------------------------------------------------------------

. test bmi male_bmi

 ( 1)  bmi = 0
 ( 2)  male_bmisqr = 0

           chi2(  2) =    5.47
         Prob > chi2 =    0.0648

. test bmisqr male_bmisqr

 ( 1)  bmisqr = 0
 ( 2)  male_bmisqr = 0

           chi2(  2) =    8.19
         Prob > chi2 =    0.0166

. test bmi bmisqr male_bmi male_bmisqr

 ( 1)  bmi = 0
 ( 2)  bmisqr = 0
 ( 3)  male_bmi = 0
 ( 4)  male_bmisqr = 0

           chi2(  4) =   15.22
         Prob > chi2 =    0.0016
```

Grade distribution

Maximum possible: 132 (4 points each problem)
Highest achieved : 126

Mean (SD) : 93.04 (18.0)

90^{th} Percentile : 117
75^{th} Percentile : 108
50^{th} Percentile : 94
25^{th} Percentile : 79