Bios 312:
Modern Regression Analysis

Final Examination
April 28, 2011

Name: **KEY**

Instructions: Please provide concise answers to all questions. Questions are of varying levels of difficult, so you may find it advantageous to skip questions you find especially difficult, and return to these questions at the end of the exam.

You are allowed up to six (6) pages of your own notes to assist you when taking the exam.

You may use a calculator to assist with arithmetic. When making intermediate calculations, always use at least four significant digits; report at least three significant digits.

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

Please adhere to the following pledge. If you are 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

_____

This exam consists of

- 11 total pages including the Appendix of Results
- There are 135 total points
  - Question 1: parts (a) – (e), 25 points
  - Question 2: parts (a) – (d), 20 points
  - Question 3: parts (a) – (m), 65 points
  - Question 4: parts (a) – (b), 10 points
  - Question 5: parts (a) – (e), 15 points

Question 1 (25 points total, 5 points each). Consider the following univariate regression models exploring the association between survival and body mass index (BMI). BMI is a continuous measure of weight divide by height-squared (BMI = kg / m^2).

For each model given below, provide an interpretation of the slope parameter in terms of which summary measure is compared across groups. Indicate whether such an analysis would be appropriate in this dataset. Note that while some subjects are censored in this dataset, no subject is censored before 4 years time. (You do not have any results for the following analyses. Just indicate what you would be examining.)

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

The slope estimates the expected difference in BMI between those subjects who die within 4 years and those subjects who survive more than 4 years. Since no subjects were censored before 4 years, this analysis is appropriate.

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

The slope estimates the difference in the probability of death comparing two subjects who differe in their BMI by 1 kg/m^2. Again, since no subjects are censored before 4 years, the analysis is appropriate. (Robust standard errors would need to be used to get proper standard error estiamtes).

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

The slope estimates the difference in the log odds of death comparing two subjects who differe in their BMI by 1 kg/m^2. Equivalently, this is the the log odds ratio comparing the two groups. Since no subject were censored before 4 years, this analysis is appropriate.

1.d) A proportional hazards regression of the observation time and a variable indicating that death occurred within 4 years (response) on BMI (predictor).

The slope estimates the difference in the log hazard ratio of death comparing two subjects who differe in their BMI by 1 kg/m^2. It will be valid approach if we have censored data or not.

1.e) Very briefly indicate the relative advantages and disadvantages of these four approaches.

The analysis based on dichotomizing death would have less statistical power than the PH regression, which measures time to event continuously. Among parts (a-c), statistical precision would likely be similar. I would choose the model based on the ease of interpretation, and/or the presumed direction of my cause-effect relationship. My belief is that BMI is causing death, so I would choose a model with death as the outcome. Futhermore, odds ratios are easier for me to interpret, but it could be argued that differences in proportions are preferred by some audiences.

Question 2 (20 points total, 5 points each). Each of following logistic regression models examine the association between some function of BMI and death occurring within 4 years. In Model 2.1, BMI is included as a linear term. In 2.2, BMI is modeled using a linear and a quadratic term. Model 2.3 uses dummy variables to indicate medium and high BMI using cut-points of 18 and 30. The reference group for model 2.3 is low BMI (defined to be BMI of 18 or lower).

More specifically, the models are given as:

Model 2.1: log(odds of death) = $\alpha_0 + \alpha_1$*bmi
Model 2.2: log(odds of death) = $\beta_0 + \beta_1$*bmi + β_2*bmi^2
Model 2.3: log(odds of death) = $\gamma_0 + \gamma_1$*bmi_middle + γ_2*bmi_high

bmi_middle = 1 if BMI is within the range (18,30], and 0 otherwise
bmi_high = 1 if BMI is above 30, and 0 otherwise

Each of these models are fit in Appendix A. For each model, output is provided for the log-odds of death (obtained by the logit command) and the odds of death (obtained by the logistic command). Use this output to answer the following questions.

2.1) Using the results from Model 2.1, is there an association between the odds of death and BMI? Justify your answer.

There is no evidence of an association. The p-value for the BMI coefficient is not significantly different from zero (p = 0.27 from Wald test; p = 0.26 from Likelihood Ratio test).

2.2) Using the results from Model 2.2, is there an association between the odds of death and BMI? Justify your answer.

There is no evidence of an association using either Likelihood ratio tes t (p = 0.07) or the Wald test (p = 0.06). The test of bmi + bmi_square is meaningless here.

2.3) Using the results from Model 2.3, is there an association between the odds of death and BMI? Justify your answer.

Modeled as a categorical variable, there appears to be an association between BMI and odds of death from either the LR test or the Wald tests (p = 0.005 and p=0.008, respectively).

2.4) Compared the results from each of the models. Based on the given output, is there any evidence to suggest that there is a non-linear association between the odds of death and BMI? Why or why not?

From Model 2.2, there is evidence that the squared term for BMI is different from zero (p = 0.034), which indicates a departure from linearity. However, overall there is no significant effect of BMI from this model. Less formally, the estimates from Model 2.3 indicate non-linearity in that low BMI is associated with high risk of death, while medium and high BMI have similar risks of death. This implies a non-linear relationship that might be capture well by splines.

Question 3(65 points total, 5 points each). Appendix B contains the results of a logistic regression performed on a variable indicating death observed with 4 years on BMI, age, gender, the BMI-gender interaction, and the BMI-age interaction. Use the output in Appendix B to answer the following questions.

3.a) What is the estimated odds of death within 4 years for a 60 year old female with a BMI of 30 kg/m^2?

log(odds) = -12.80 + .1182*60 + 30*.0474 = -4.286; odds = exp(-4.286) = 0.01376

3.b) What is the estimated probability of death within 4 years for a 60 year old female with a BMI of 30 kg/m^2?

prob = odds / (1 + odds) = .01376 / 1.01376 = 0.01357

3.c) What is the estimated odds of death within 4 years for a 61 year old female with a BMI of 30 kg/m^2?

0.01376 * 1.125 = 0.01548

3.d) What is the estimated odds of death within 4 years for a 60 year old male with a BMI of 30 kg/m^2?

log(odds) = -12.80 + .1182*60 + 30*.0474 +7.398 – 0.0981*30 - .0515*60 = -2.921
odds = exp(-2.921) = 0.0539

3.e) What is the interpretation of the intercept in the regression model obtained using the logit command?

The log(odds of death) in a female subject with a BMI of 0 and age of 0. Since BMI is weight divided by height-squared, the subject must have be a newborn with no weight but some height.

3.f) What is the interpretation of the slope parameter for BMI obtained using the logit command?

The log odds ratio comparing groups of females who differ in BMI by 1 kg/m^2, but have the same age.

3.g) What is the interpretation of the slope parameter for age obtained using the logit command?

The log odds ratio comparing groups of females who differ in age by 1 year, but have identical BMI.

3.h) What is the interpretation of the slope parameter for gender obtained using the logit command?

The log odds ratio comparing a newborn male with no height but some weight to a newborn female with no height but some weight.

3.i) What is the interpretation of the slope parameter for the gender-BMI interaction obtained using the logit command?

The difference in the log odds ratio comparing groups of males have the same age but differing in BMI by 1 kg/m^2 to the log odds ratio comparing groups offe males have the same age but differing in BMI by 1 kg/m^2.

3.j) Using this model, explain how you would test if BMI is significantly associated with death within 4 years. Give the null and alternative hypothesis for this test.

We would need to simulatenously test all coefficients that include BMI are equal to zero (the null hypthesis). The would include the bmi and the bmi-gender interaction. The alternaitve hypothesis is that at least one of these two coefficients is not equal to zero.

3.k) Using this model, explain how you would test if gender is significantly associated with death within 4 years. Give the null and alternative hypothesis for this test.

We need to simulatenously test all coefficients that include gener (male) are equal to zero (the null hypthesis). The would include the gender main effect and the interaction of gender with BMI and gender with age. The alternaitve hypothesis is that at least one of these three coefficients is not equal to zero.

3.l) Suppose we fit a logistic regression model of death with 4 years on BMI and age in just females. What would the estimate of the intercept, BMI slope, and age slope in such a model?

These would just be the results given directly from the model, setting all male terms equal to zero: -12.80 + .1182*age + .0474*bmi

3.m) Suppose we fit a logistic regression model of death with 4 years on BMI and age in just males. What would the estimate of the intercept, BMI slope, and age slope in such a model?

We now need to add the effect of gender interactions for each coefficient
(-12.80 + 7.40) + (.1182 – 0.0515)*age + (.0474 – 0.0981)*bmi
-5.4 + 0.0667*age – 0.0507*bmi

Question 4 (20 points total, 10 points each). Suppose we are interested in exploring the association between systolic blood pressure and age and gender. Consider two possible study designs:

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

- Study B: Gather five measurements made one year apart on each of 1,000 independent subjects of both sexes between the age of 60 and 80

4.a) Is Study A or Study B more likely to provide more statistical precision to assess an association between blood pressure and age? Explain why.

Study B. Both studies have the same number of total observations, but study B takes repeated measurements on the same subject over time. Blood pressure measurements on the same subject are likely to be positively correlated. When looking at changes within an invidiual over age, we will be able to make comparisons of blood pressure within the same subject. Hence, we will be looking at differences of positively correlated observation, which tend to have smaller standard errors than differences between independent observations. The degree of improvement in the precision will depend on the magnitude of the correlation and the varaibility of the ages that will result from the two study designs. I am assuming the variability of the ages is similar in the two designs, and the correlation is over 0.5.

4.b) Is Study A or Study B more likely to provide more statistical precision to assess an association between blood pressure and gender? Explain why.

Study A. Both designs have the same number of total observations, and blood pressure measurements are likely to be positively correlated within individuals. The primary comparison of interest here, gender, can only be measured among different subjects. Any repeated measures on the same subject will be measurements in the same gender group. Hence, we will be using sums of positively correlated observations, which tends to lead to larger standard errors than sums of the same number of independent observations.

Question 5 (15 points total, 3 points each). The Scholastic Aptitude Test (SAT) is a standardized test that many colleges and universities use in evaluating undergraduate students for admission. Assume that the SAT is rigorously designed and evaluated so that, each year, scores follow a Normal distribution with a mean of 500 points and standard deviation of 100 points. An investigator has access to a random sample of SAT scores from North Carolina and Tennessee. For each score, the investigator also has data on whether the test was taken in the junior or senior year of high school. They want to evaluate the impact of state and year on SAT scores using the following linear regression model:

$$E[SAT \mid state, year] = \beta_0 + \beta_1 * state + \beta_2 * year + \beta_3 * state * year$$

where state=1 if Tennessee and state=0 if North Carolina,
year=1 if taken in senior year or year=0 if taken in junior year

For each of the following assumptions of classical linear regression, indicate whether or not you expect the assumption to hold for this analysis, and if this assumption is necessary to make statistical inference about association, means in groups, and/or prediction (forecasting) of new individual observations. If you believe an assumption might not hold, briefly explain how you would evaluate the assumption.

5.1 All of the observations are independent
Assumption needed for all inference types. If the data were collected in one calendar year, then all of the observations should be independent. If the data were collected over multiple years, then some subjects might have taken the test as both a junior and senior. Evaluating these assumptions would require detailed knowledge of the study design.

5.2 The parameter estimates (the βs) follow a Normal distribution
Assumptions need for all types of inference. It seems reasonable that the data, conditional on state and year, should be Normally distributed. Thus, the parameter estimates will be Normally distributed (even if there are few observations).

5.3 Constant variance across groups (homoskedasticity)
Assumptions need for all types of inference, unless using robust standard errors. Robust standard erros helps for associations and means, but not prediction. With Normal data, this assumption seems plausible. It could be checked by plots of the residuals versus the fitted values or other diagnostic measures.

5.4 The mean model has been appropriately specified
Needed for means in groups and predictions. Here, the mean model must be appropriately specified because we are precisely modeling the mean for each possible level of state and year by using the interaction term.

5.5 The residuals follow a Normal distribution
Also seems like this is a reasonable assumption because the data are Normally distributed. Could be evaluated by looking at a histogram, qq-plot, or similar distribution plot of the residuals. Needed for predcitions only.

Appendix for Question 2: Different dose-response models for BMI (Appendix A)

```
. gen death4 = 0
. replace death4 = 1 if obstime <=4 & death==1
(238 real changes made)


. gen bmi_middle = 0
. replace bmi_middle = 1 if bmi <=30 & bmi>18
(1983 real changes made)

. gen bmi_high = 0
. replace bmi_high = 1 if bmi >30
(495 real changes made)

. gen bmi_sqr = bmi*bmi
```

Model 2.1: BMI modeled as a linear function

```
. logit death4 bmi

Logistic regression                                     Number of obs   =       2500
                                                        LR chi2(1)      =       1.26
                                                        Prob > chi2     =     0.2622
Log likelihood = -785.38747                             Pseudo R2       =     0.0008

------------------------------------------------------------------------------
      death4 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |  -.0162804   .0146713    -1.11   0.267    -.0450356    .0124747
       _cons |   -1.81808    .394575    -4.61   0.000    -2.591432   -1.044727
------------------------------------------------------------------------------

. logistic death4 bmi

Logistic regression                                     Number of obs   =       2500
                                                        LR chi2(1)      =       1.26
                                                        Prob > chi2     =     0.2622
Log likelihood = -785.38747                             Pseudo R2       =     0.0008

------------------------------------------------------------------------------
      death4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |   .9838514   .0144344    -1.11   0.267     .9559634    1.012553
------------------------------------------------------------------------------
```

Model 2.2: BMI modeled as a quadratic function

```
.
. logit death4 bmi bmi_sqr


Logistic regression                               Number of obs   =       2500
                                                  LR chi2(2)      =       5.24
                                                  Prob > chi2     =     0.0727
Log likelihood = -783.39493                       Pseudo R2       =     0.0033

------------------------------------------------------------------------------
      death4 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |  -.2007815   .0889987    -2.26   0.024    -.3752158   -.0263472
     bmi_sqr |   .0031772   .0014965     2.12   0.034     .0002441    .0061103
       _cons |   .7648577   1.293666     0.59   0.554    -1.770681    3.300397
------------------------------------------------------------------------------

. logistic death4 bmi bmi_sqr

Logistic regression                               Number of obs   =       2500
                                                  LR chi2(2)      =       5.24
                                                  Prob > chi2     =     0.0727
Log likelihood = -783.39493                       Pseudo R2       =     0.0033

------------------------------------------------------------------------------
      death4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |   .8180912   .0728091    -2.26   0.024      .687141    .9739969
     bmi_sqr |   1.003182   .0015013     2.12   0.034     1.000244    1.006129
------------------------------------------------------------------------------

. test bmi+bmi_sqr=0

 ( 1)  [death4]bmi + [death4]bmi_sqr = 0

           chi2(  1) =    5.10
         Prob > chi2 =    0.0240

. test (bmi=0) (bmi_sqr=0)

 ( 1)  [death4]bmi = 0
 ( 2)  [death4]bmi_sqr = 0

           chi2(  2) =    5.59
         Prob > chi2 =    0.0611
```

Model 2.3: BMI modeled using indicator variables

```
.
. logit death4 bmi_middle bmi_high

Logistic regression                                     Number of obs   =       2500
                                                        LR chi2(2)      =       8.44
                                                        Prob > chi2     =     0.0147
Log likelihood = -781.79504                             Pseudo R2       =     0.0054

------------------------------------------------------------------------------
      death4 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  bmi_middle |  -1.235024   .4846175    -2.55   0.011    -2.184857   -.2851912
    bmi_high |  -1.535122   .5083081    -3.02   0.003    -2.531388   -.5388565
       _cons |  -.9808292   .4787136    -2.05   0.040    -1.919091   -.0425679
------------------------------------------------------------------------------

. logistic death4 bmi_middle bmi_high

Logistic regression                                     Number of obs   =       2500
                                                        LR chi2(2)      =       8.44
                                                        Prob > chi2     =     0.0147
Log likelihood = -781.79504                             Pseudo R2       =     0.0054

------------------------------------------------------------------------------
      death4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  bmi_middle |   .2908277   .1409402    -2.55   0.011     .1124938    .7518705
    bmi_high |   .2154294   .1095045    -3.02   0.003     .0795486     .583415
------------------------------------------------------------------------------

.
. test bmi_middle+bmi_high=0

 ( 1)  [death4]bmi_middle + [death4]bmi_high = 0

           chi2(  1) =    8.06
         Prob > chi2 =    0.0045

. test (bmi_middle=0) (bmi_high=0)

 ( 1)  [death4]bmi_middle = 0
 ( 2)  [death4]bmi_high = 0

           chi2(  2) =    9.63
         Prob > chi2 =    0.0081
```

Appendix for Question 3 (Appendix B)

```
. gen male_bmi = male*bmi
. gen male_age = male*age

. logit death4 age bmi male male_bmi male_age


Logistic regression                                     Number of obs   =       2500
                                                        LR chi2(5)      =     121.78
                                                        Prob > chi2     =     0.0000
Log likelihood = -725.12598                             Pseudo R2       =     0.0775

------------------------------------------------------------------------------
      death4 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .1181702   .0179938     6.57   0.000      .082903    .1534374
         bmi |   .0474373   .0199491     2.38   0.017     .0083377    .0865368
        male |   7.397875   2.092839     3.53   0.000     3.295986    11.49976
    male_bmi |  -.0980824   .0324695    -3.02   0.003    -.1617214   -.0344434
    male_age |  -.0515358   .0231121    -2.23   0.026    -.0968347   -.0062369
       _cons |  -12.80381   1.564693    -8.18   0.000    -15.87056   -9.737071
------------------------------------------------------------------------------

. logistic death4 age bmi male male_bmi male_age

Logistic regression                                     Number of obs   =       2500
                                                        LR chi2(5)      =     121.78
                                                        Prob > chi2     =     0.0000
Log likelihood = -725.12598                             Pseudo R2       =     0.0775

------------------------------------------------------------------------------
      death4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.125436   .0202509     6.57   0.000     1.086436    1.165835
         bmi |    1.04858   .0209183     2.38   0.017     1.008373    1.090392
        male |   1632.511   3416.584     3.53   0.000     27.00403    98692.45
    male_bmi |   .9065742    .029436    -3.02   0.003     .8506782    .9661431
    male_age |   .9497697   .0219512    -2.23   0.026     .9077061    .9937825
------------------------------------------------------------------------------
```