Biost 518
Applied Biostatistics II

Midterm Examination Key
February 8, 2006

Name: _____ Disc Sect: M W F

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. 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.

1. **(For all calculations in this problem, please use at least 4 significant digits.)** Suppose we are interested in the association between systolic blood pressure, age, and weight. The following are the results of a linear regression analysis of data on 735 elderly Americans. The variable definitions are
   - *sbp:* systolic blood pressure in mm Hg
   - *age:* age in years
   - *weight:* weight in pounds

```
. regress sbp age weight, robust

Linear regression                               Number of obs =     735
                                                F(  2,   732) =    5.90
                                                Prob > F      =  0.0029
                                                R-squared     =  0.0159
                                                Root MSE      =  19.533
------------------------------------------------------------------------------
             |               Robust
         sbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .4669811   .1362026     3.43   0.001     .1995868    .7343754
      weight |   .0264977   .0245915     1.08   0.282    -.0217805    .0747759
       _cons |   92.04612   11.79386     7.80   0.000     68.89229    115.1999
------------------------------------------------------------------------------
```

   a. (5 points) Based on the above regression model, what is the best estimate for the mean systolic blood pressure in 70 year old subjects who weigh 150 pounds?

   Ans: 92.05 + 70 * 0.4670 + 150 * 0.02650 = 128.7

b. (5 points) Based on the above regression model, what is the best estimate for the mean systolic blood pressure in 71 year old subjects who weigh 150 pounds?

Ans: 92.05 + 71 * 0.4670 + 150 * 0.02650 = 128.7 + 0.4670 = 129.2

c. (5 points) Based on the above regression model, what is the best estimate for the mean systolic blood pressure in 70 year old subjects who weigh 160 pounds?

Ans: 92.05 + 70 * 0.4670 + 160 * 0.02650 = 128.7 + 0.2650 = 128.98

d. (5 points) Based on the above regression model, what is the best estimate for the difference in mean systolic blood pressure between 70 year old subjects who weigh 150 pounds and 71 year old subjects who weigh 150 pounds? Provide a confidence interval for this difference.

Ans: 0.4670; 95% CI 0.1996 to 0.7344 *(This is merely asking about the slope for age.)*

e. (5 points) Based on the above regression model, what is the best estimate for the difference in mean systolic blood pressure between 70 year old subjects who weigh 160 pounds and 71 year old subjects who weigh 160 pounds? Provide a confidence interval for this difference.

Ans: 0.4670; 95% CI 0.1996 to 0.7344 *(This is still merely asking about the slope for age, because there is no interaction in the model.)*

f. (5 points) Based on the above regression model, what is the best estimate for the difference in mean systolic blood pressure between 70 year old subjects who weigh 150 pounds and 70 year old subjects who weigh 151 pounds? Provide a confidence interval for this difference.

Ans: 0.02650; 95% CI -0.02178 to 0.07478 *(This is merely asking about the slope for weight.)*

g. (5 points) Based on the above regression model, what is the best estimate for the difference in mean systolic blood pressure between 70 year old subjects who weigh 150 pounds and 70 year old subjects who weigh 160 pounds? Provide a confidence interval for this difference.

Ans: 0.2650; 95% CI -0.2178 to 0.7478 *(This is merely asking about 10 times the slope for weight. We can just multiply the limits of the confidence interval by 10 as well.)*

h. (5 points) Provide an interpretation for the intercept in the above regression model. What scientific use would you make of this estimate?

Ans: The estimated average SBP for newborns who weigh 0 pounds is 92.05 mm Hg. There are no such people, so there is no scientific use of this.

i. (5 points) Provide an interpretation for the slope for the age predictor in the above regression model. What scientific use would you make of this estimate?

Ans: The mean SBP is estimated to average 0.4670 mm Hg higher for every year difference in age between two groups having the same weight, with the older group tending toward the higher SBP. This can be used to describe a first order trend in the association between mean SBP and age after adjusting for weight. If the association for both age and weight followed a straight line relationship, I might also use this to predict the mean SBP in every age and weight group.

j. (5 points) Is there evidence that the slope for the age predictor is different from 0? State your evidence.

Ans: Yes, the t test for the age slope parameter is highly statistically significant: P = 0.001.

k. (5 points) Is there evidence of an association between systolic blood pressure and age after adjustment for weight? State your evidence.

Ans: Yes, because the t test for the age slope parameter is highly statistically significant: P = 0.001.

l. (5 points) Provide an interpretation for the slope for the weight predictor in the above regression model. What scientific use would you make of this estimate?

Ans: The mean SBP is estimated to average 0.02650 mm Hg higher for every pound difference in weight between two groups having the same age, with the heavier group tending toward the higher SBP. This can be used to describe a first order trend in the association between mean SBP and weight after adjusting for age. If the association for both age and weight followed a straight line relationship, I might also use this to predict the mean SBP in every age and weight group.

m. (5 points) Is there evidence that the slope for the weight predictor is different from 0? State your evidence.

Ans: No, the t test for the weight slope parameter is not statistically significant: P = 0.282.

n. (5 points) Is there evidence of an association between systolic blood pressure and weight after adjustment for age? State your evidence.

Ans: No, because the t test for the weight slope parameter is not statistically significant: P = 0.282.

2. The following analysis added second order terms for both age and weight to the above regression model (so a term for age squared and weight squared).

```
. g agesqr= age * age
. g wtsqr= weight * weight
. regress sbp age weight agesqr wtsqr, robust
Linear regression                               Number of obs =     735
                                                F(  4,   730) =    3.03
                                                Prob > F      =  0.0171
                                                R-squared     =  0.0161
                                                Root MSE      =  19.558
------------------------------------------------------------------------------
             |               Robust
         sbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.3235518   2.411916    -0.13   0.893     -5.05867    4.411567
      weight |  -.0048617   .1898531    -0.03   0.980     -.377585    .3678615
      agesqr |   .0051194   .0155819     0.33   0.743    -.0254712      .03571
       wtsqr |   .0000935   .0005468     0.17   0.864      -.00098    .0011671
       _cons |   124.9112   92.87914     1.34   0.179    -57.43086    307.2533
------------------------------------------------------------------------------

. test agesqr wtsqr
 ( 1)  agesqr = 0
 ( 2)  wtsqr = 0
       F(  2,   730) =    0.08
            Prob > F =    0.9277
```

a. (5 points) Is there statistical evidence that either the age effect on mean systolic blood pressure or the weight effect on mean systolic blood pressure is nonlinear? State your evidence.

Ans: Based on the "multiple partial F test" that both the agesqr and wtsqr terms are 0, the lack of statistical significance (P = 0.9277) suggests that we do not have evidence that either the association between mean SBP and weight after adjustment for age or the association between mean SBP and age after adjustment for weight are markedly nonlinear.

b. (5 points) Based on your answer to part a, how reliable do you think the estimates you provided in problem 1, parts a-c are?

Ans: We cannot prove exact linearity without an infinite sample size, but given the lack of evidence for a markedly nonlinear trend in this sample of 735, I tend to regard the estimates of mean SBP in each age and weight group relatively reliable. *(Ideally, I would consider the limits of the CI for agesqr and wtsqr and see how far those curves depart from a straight line over the ranges of ages and weights sampled.)*

3. **(For all calculations in this problem, please use at least 4 significant digits.)** Below are results of analyses on 735 elderly Americans comparing rates of high blood pressure (systolic blood pressure greater than 160 mmHg) across racial groups. The following variables are used:
   - *sbp160*: an indicator that the subject's systolic blood pressure is greater than 160 mmHg (0= no, 1= yes)
   - *race*: a coded variable indicating the subject's race (1= Caucasian, 2= African American, 3= Asian American)
   - *cauc*: an indicator that the subject's race is Caucasian (0= no, 1=yes)
   - *afram*: an indicator that the subject's race is African American (0= no, 1=yes)
   - *asianam*: an indicator that the subject's race is Asian American (0= no, 1=yes)

   The following are the results from descriptive statistics and four alternative logistic regression models. Note that for each regression model, I provide results using the Stata command "logit".

   Tabulated frequencies of high blood pressure by race (counts and row percentages):

```
. tabulate race sbp160, row
```

race	sbp160 0	sbp160 1	Total
1	537	35	572
	93.88%	6.12%	100.00
2	90	14	104
	86.54%	13.46%	100.00
3	54	5	59
	91.53%	8.47%	100.00
Total	681	54	735
	92.65%	7.35%	100.00

   Model A: Logistic regression of *sbp160* on *race*:

```
. logit sbp160 race, robust

Logistic regression                               Number of obs   =        735
                                                  Wald chi2(1)    =       3.60
                                                  Prob > chi2     =     0.0578
Log pseudolikelihood = -191.53039                 Pseudo R2       =     0.0074
------------------------------------------------------------------------------
             |               Robust
      sbp160 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        race |   .3527991   .1859694     1.90   0.058    -.0116942    .7172924
       _cons |  -3.015406   .3000273   -10.05   0.000    -3.603449   -2.427364
------------------------------------------------------------------------------
```

Model B: Logistic regression of *sbp160* on *cauc* and *afram*:

```
. logit sbp160 cauc afram, robust

Logistic regression                               Number of obs   =        735
                                                  Wald chi2(2)    =       6.75
                                                  Prob > chi2     =     0.0343
Log pseudolikelihood = -189.89871                 Pseudo R2       =     0.0158
------------------------------------------------------------------------------
             |               Robust
     sbp160 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+----------------------------------------------------------------
       cauc |  -.3511039   .4992908    -0.70   0.482    -1.329696    .6274881
      afram |   .5187938   .5490613     0.94   0.345    -.5573467    1.594934
      _cons |  -2.379546    .467778    -5.09   0.000    -3.296374   -1.462718
------------------------------------------------------------------------------

. test cauc afram
 ( 1)  cauc = 0
 ( 2)  afram = 0
           chi2(  2) =    6.75
         Prob > chi2 =    0.0343
```

Model C: Logistic regression of *sbp160* on *cauc* and *asianam*:

```
. logit sbp160 cauc asianam, robust

Logistic regression                               Number of obs   =        735
                                                  Wald chi2(2)    =       6.75
                                                  Prob > chi2     =     0.0343
Log pseudolikelihood = -189.89871                 Pseudo R2       =     0.0158
------------------------------------------------------------------------------
             |               Robust
     sbp160 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+----------------------------------------------------------------
       cauc |  -.8698977   .3363439    -2.59   0.010     -1.52912   -.2106758
   asianam  |  -.5187938   .5490613    -0.94   0.345    -1.594934    .5573467
      _cons |  -1.860752   .2874928    -6.47   0.000    -2.424228   -1.297277
------------------------------------------------------------------------------

. test cauc asianam
 ( 1)  cauc = 0
 ( 2)  asianam = 0
           chi2(  2) =    6.75
         Prob > chi2 =    0.0343
```

Model D: Logistic regression of *sbp160* on *afram* and *asianam*:

```
. logit sbp160 afram asianam, robust

Logistic regression                               Number of obs   =        735
                                                  Wald chi2(2)    =       6.75
                                                  Prob > chi2     =     0.0343
Log pseudolikelihood = -189.89871                 Pseudo R2       =     0.0158
------------------------------------------------------------------------------
             |               Robust
     sbp160 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+----------------------------------------------------------------
      afram |   .8698977   .3363439     2.59   0.010     .2106758     1.52912
   asianam  |   .3511039   .4992908     0.70   0.482    -.6274881    1.329696
      _cons |   -2.73065   .1745712   -15.64   0.000    -3.072803   -2.388497
------------------------------------------------------------------------------
```

a. (5 points) Using the proportion of each race having high blood pressure (greater than 160 mm Hg) as derived from the crosstabulation of *sbp160* and *race*, what are the odds of high blood pressure for each race (Caucasian, African American, and Asian American)?

Ans: Sample odds of hypertension (HTN) (defined as SBP > 160) by race:

Caucasian: 0.0612 / (1 – 0.0612) = 0.06519

African American: 0.1346 / (1 – 0.1346) = 0.1555

Asian American: 0.0847 / (1 – 0.0847) = 0.09254

b. (5 points) Using Model A, what is the estimated odds that a Caucasian would have a systolic blood pressure greater than 160 mm Hg? What is the estimated probability that a Caucasian would have a systolic blood pressure greater than 160 mm Hg?

Ans: Caucasians are coded as *race=1*, so

estimated log odds of HTN: -3.015 + 1 * 0.3528 = -2.662

estimated odds of HTN: exp (-2.662) = 0.06981

estimated probability of HTN: 0.06981 / (1 + 0.06981) = 0.06525

c. (5 points) Using Model A, what is the estimated odds that an African American would have a systolic blood pressure greater than 160 mm Hg? What is the estimated probability that an African American would have a systolic blood pressure greater than 160 mm Hg?

Ans: African Americans are coded as *race=2*, so

estimated log odds of HTN: -3.015 + 2 * 0.3528 = -2.309

estimated odds of HTN: exp (-2.309) = 0.09936

estimated probability of HTN: 0.09936 / (1 + 0.09936) = 0.09038

d. (5 points) Using Model A, what is the estimated odds that an Asian American would have a systolic blood pressure greater than 160 mm Hg? What is the estimated probability that an Asian American would have a systolic blood pressure greater than 160 mm Hg?

Ans: Asian Americans are coded as *race=3*, so

estimated log odds of HTN: -3.015 + 3 * 0.3528 = -1.957

estimated odds of HTN: exp (-1.957) = 0.1413

estimated probability of HTN: 0.1413 / (1 + 0.1413) = 0.1238

e. (5 points) How do the estimates derived from Model A compare with the proportions reported with the crosstabulation of *sbp160* and *race* reported in the descriptive statistics? Briefly explain why they might agree or disagree.

Ans: They do not agree. Fitting race as a continuous variable means that we are borrowing information across groups in order to estimate the log odds (and odds and probabilities) of HTN for each race group. *(It is, of course, inappropriate to borrow information across unordered categories in this fashion.)*

f. (5 points) Using Model B, what is the estimated odds that a Caucasian would have a systolic blood pressure greater than 160 mm Hg? What is the estimated probability that a Caucasian would have a systolic blood pressure greater than 160 mm Hg?

Ans: In model B, we are fitting a nominal variable with 3 groups using two dummy variables and no other covariates. The estimated odds and probability will agree exactly with the estimates based on the sample descriptive statistics: estimated odds of HTN are 0.06519, estimated probability of HTN is 0.0612.

(Of course, if you didn't recognize the easy approach to answer this problem, you could work out the formulas:
Caucasians are coded as cauc=1 and afram=0, so
*estimated log odds of HTN: -2.380 + 1 * (-0.3511) + 0 * 0.5188 = -2.731*
estimated odds of HTN: exp (-2.731) = 0.06515
estimated probability of HTN: 0. 06515 / (1 + 0. 06515) = 0.06117
which agree with part a up to the three significant digits given in the descriptive statistics.)

g. (5 points) Using Model B, what is the estimated odds that an African American would have a systolic blood pressure greater than 160 mm Hg? What is the estimated probability that an African American would have a systolic blood pressure greater than 160 mm Hg?

Ans: Again, the estimated odds and probability will agree exactly with the estimates based on the sample descriptive statistics: estimated odds of HTN are 0.1555, estimated probability of HTN is 0.1346.

(Of course, if you didn't recognize the easy approach to answer this problem, you could work out the formulas:
African Americans are coded as cauc=0 and afram=1, so
*estimated log odds of HTN: -2.380 + 0 * (-0.3511) + 1 * 0.5188 = -1.861*
estimated odds of HTN: exp (-1.861) = 0.1555
estimated probability of HTN: 0. 1555 / (1 + 0. 1555) = 0.1346
which agree with part a.)

h. (5 points) Using Model B, what is the estimated odds that an Asian American would have a systolic blood pressure greater than 160 mm Hg? What is the estimated probability that an Asian American would have a systolic blood pressure greater than 160 mm Hg?

Ans: Again, the estimated odds and probability will agree exactly with the estimates based on the sample descriptive statistics: estimated odds of HTN are 0.09254, estimated probability of HTN is 0.0847.

(Of course, if you didn't recognize the easy approach to answer this problem, you could work out the formulas:
Asian Americans are coded as cauc=0 and afram=0, so
*estimated log odds of HTN: -2.380 + 0 * (-0.3511) + 0 * 0.5188 = -2.380*
estimated odds of HTN: exp (-2.380) = 0.09255
estimated probability of HTN: 0. 09255 / (1 + 0. 09255) = 0.08471
which agree with part a up to the three significant digits given in the descriptive statistics.)

i. (5 points) How do the estimates derived from Model B compare with the proportions reported with the crosstabulation of *sbp160* and *race* reported in the descriptive statistics? Briefly explain why they might agree or disagree.

Ans: They agree exactly within roundoff error, because we are fitting three groups with two dummy variables and no other covariates. *(In this setting, we are not borrowing information across groups in any way, so the estimates have to agree.)*

j. (10 points) Using Models C and D, what would be the estimated odds and probabilities that a member of each race group would have a systolic blood pressure greater than 160 mm Hg? Briefly explain how you derived your answer.

Ans: They would agree exactly (within roundoff error) with the results for Model B and the sample descriptive statistics, because we are fitting three groups with two dummy variables and no other covariates. *(In this setting, we are not borrowing information across groups in any way, so the estimates have to agree. Of course, you could also have worked out all of the cases, just as given above for Model B.)*

k. (5 points) Which of the four models would you use to address the question of an association between the prevalence of high blood pressure and race? Why?

Ans: Model A is bad, because it treats an unordered categorical variable as if it were linear continuous. The other three models are just different parameterizations of the exact same predictive model, and they are the appropriate way to model a nominal variable.

l. (5 points) Is there a statistically significant association between high blood pressure and race? Provide the P value you use to answer this question.

Ans: Because in the appropriate models (either B, C, or D), race is modeled with two predictors, we must use the "multiple partial" test considering that both dummy variable slope parameters are 0. Because P = 0.0343, we can reject the null hypothesis of no association between the odds of hypertension and race.

4. **(For all calculations in this problem, please use at least 4 significant digits.)** I asked a research assistant to investigate whether there was statistical evidence that sex modified the association between systolic blood pressure and age. I had, of course, expected the student to perform a linear regression of *sbp* (systolic blood pressure in mm Hg) including terms for age (variable *age* measured in years), an indicator of male sex (variable *male*=0 for females, *male*=1 for males), and a variable *maleage= male * age*. Unfortunately, the research assistant provided me with the following two subgroup analyses: linear regressions of *sbp* on *age* within each sex group separately.

```
Linear regression model for females:
. regress sbp age if male==0, robust

Linear regression                               Number of obs =     369
                                                F(  1,   367) =   15.79
                                                Prob > F      =  0.0001
                                                R-squared     =  0.0377
                                                Root MSE      =  19.618
------------------------------------------------------------------------------
             |               Robust
         sbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .7371504   .1854805     3.97   0.000     .3724125    1.101888
       _cons |   76.46756    13.7411     5.56   0.000     49.44639    103.4887
------------------------------------------------------------------------------

Linear regression model for males:
. regress sbp age if male==1, robust

Linear regression                               Number of obs =     366
                                                F(  1,   364) =    0.78
                                                Prob > F      =  0.3777
                                                R-squared     =  0.0023
                                                Root MSE      =  19.377
------------------------------------------------------------------------------
             |               Robust
         sbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .1662855   .1882744     0.88   0.378    -.2039566    .5365275
       _cons |   118.4666   14.02984     8.44   0.000      90.8769    146.0563
------------------------------------------------------------------------------
```

   a. (10 points) Is there a statistically significant difference between the age slope for females and the age slope for males?

Ans: Because the regression parameter estimates are approximately normally distributed and independent, we can estimate the standard error for the estimated difference by the square root of the sum of the squared standard errors:

Estimated difference in slopes (females – males): 0.7372 – 0.1663 = 0.5709

Estimated SE for estimated difference: sqrt($0.1855^2 + 0.1883^2$) = 0.2643

Z statistic for testing difference is 0: 0.5709 / 0.2643 = 2.160

Because the sample size is relatively large, we can regard the standard normal approximation for the Z statistic as relatively accurate. Thus because 2.160 is greater (in absolute value) than 1.96 (the critical value for a two-sided level 0.05 test) we regard there is a statistically significant difference between the age slope for females and that for males (P < 0.05). We therefore can reject the null hypothesis of no effect modification by sex on the association between mean SBP and age.

b. (5 points) Supposing the research assistant had fit the correct model including terms for *age, male,* and *maleage*, what would have been the estimated intercept?

Ans: Had the research assistant fit the right model, we would have estimated enough parameters to fit separate lines for females and males (i.e., each sex would have had their own intercept and slope). Hence, the estimates from the full interaction model would have corresponded to the estimates from the two subgroup analyses. The intercept in the interaction model would have corresponded to the group with *age=0, male=0,* and *maleage=0* (so newborn females), which is the interpretation of the intercept in the subgroup model for the females: 76.47.

c. (5 points) Supposing the research assistant had fit the correct model including terms for *age, male,* and *maleage*, what would have been the estimated slope for *age*?

Ans: The slope for *age* in the interaction model would have corresponded to the group with *male=0* and *maleage=0* (females are the only group for which we can contrast different values of *age* without the value of *maleage* also being different), which is the interpretation of the *age* slope for the female subgroup: 0.7372.

d. (5 points) Supposing the research assistant had fit the correct model including terms for *age, male,* and *maleage*, what would have been the estimated slope for *male*?

Ans: The slope for *male* in the interaction model would have corresponded to the group with *age=0* and *maleage=0* (newborns are the only group for which we can contrast different values of *male* without the value of *maleage* also being different), which is the interpretation of the difference between the intercept for the male subgroup and the intercept for the female subgroup: 118.5 - 76.47 = 42.03.

e. (5 points) Supposing the research assistant had fit the correct model including terms for *age, male,* and *maleage*, what would have been the estimated slope for male*age*?

Ans: The interpretation for the slope for *maleage* in the interaction model is the difference between the *age* slope for the male subgroup and the *age* slope for the female subgroup: 0.1663 – 0.7372 = - 0.5709.