Biost 518
Applied Biostatistics II

Midterm Examination Key
February 11, 2008

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.

NOTE: When you need to make calculations, always use at least four significant digits in your intermediate calculations, and report at least three significant digits. (Example: 1.045 and 0.0001234 and 1234000 each have four significant digits.)

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:_____

1. Suppose we are interested in the association between serum bilirubin, serum albumin, and the presence of edema (swelling) in a sample of patients with primary biliary cirrhosis. The following are the results of a linear regression analyses using the following variables
   - *bili:* serum bilirubin in mg/dl
   - *albumin:* serum albumin in g/dl
   - *edema:* indicator of edema (0= no, 1= yes)

```
. tabstat albumin bili, stat(n mean sd min q max) col(stat) by(edema)
Summary for variables: albumin bili
     by categories of: edema
```

vaiable	N	mean	sd	min	p25	p50	p75	max
→ edema == 0								
albumin	77	3.623	.3852	2.54	3.4	3.65	3.87	4.64
bili	77	2.439	3.078	.3	.8	1.2	3.2	20
→ edema == 1								
albumin	23	3.098	.4204	2.27	2.74	3.13	3.41	4.06
bili	23	8.643	7.528	.6	1.4	6.6	17.1	22.5
→ all patients								
albumin	100	3.502	.4501	2.27	3.205	3.535	3.79	4.64
bili	100	3.866	5.172	.3	.8	1.5	4.6	22.5

```
. regress bili albumin

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =   22.05
       Model |   486.49878     1   486.49878           Prob > F      =  0.0000
    Residual |  2162.02563    98   22.061486           R-squared     =  0.1837
-------------+------------------------------           Adj R-squared =  0.1754
       Total |  2648.52441    99  26.7527718           Root MSE      =   4.697

------------------------------------------------------------------------------
        bili |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     albumin |  -4.925514   1.048885    -4.70   0.000    -7.006992   -2.844035
       _cons |   21.11663   3.703417     5.70   0.000     13.76732    28.46594
------------------------------------------------------------------------------

. regress bili albumin edema

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  2,    97) =   20.72
       Model |  792.879897     2  396.439949           Prob > F      =  0.0000
    Residual |  1855.64451    97  19.1303558           R-squared     =  0.2994
-------------+------------------------------           Adj R-squared =  0.2849
       Total |  2648.52441    99  26.7527718           Root MSE      =  4.3738

------------------------------------------------------------------------------
        bili |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     albumin |  -2.706747   1.123111    -2.41   0.018     -4.93581    -.477683
       edema |   4.782688   1.195095     4.00   0.000     2.410755    7.154622
       _cons |   12.24582   4.099575     2.99   0.004     4.109298    20.38234
------------------------------------------------------------------------------
```

a. (5 points) Based on the regression model including both albumin and edema, what is the best estimate for the mean bilirubin in subjects with an albumin of 4.0 g/dl and no edema?

Ans: $12.25 + 0 \times 4.783 - 4 \times 2.707 = 1.419$

b. (5 points) Based on the regression model including both albumin and edema, what is the best estimate for the mean bilirubin in subjects with an albumin of 3.0 g/dl and no edema?

Ans: $12.25 + 0 \times 4.783 - 3 \times 2.707 = 4.126$

c. (5 points) Based on the regression model including both albumin and edema, what is the best estimate for the mean bilirubin in subjects with an albumin of 2.5 g/dl and no edema?

Ans: $12.25 + 0 \times 4.783 - 2.5 \times 2.707 = 5.479$

d. (5 points) Based on the regression model including both albumin and edema, what is the best estimate for the mean bilirubin in subjects with an albumin of 4.0 g/dl and edema present?

Ans: $12.25 + 1 \times 4.783 - 4 \times 2.707 = 6.202$

e. (5 points) Based on the regression model including both albumin and edema, what is the best estimate for the difference in mean bilirubin between subjects with an albumin of 4.0 g/dl and no edema and subjects with an albumin of 3.0 g/dl and no edema?

Ans: **2.707 (higher in low albumin)** *(This is just the slope for the albumin covariate)*

f. (5 points) Based on the regression model including both albumin and edema, what is the best estimate for the difference in mean bilirubin between subjects with an albumin of 4.0 g/dl and edema present and subjects with an albumin of 4.0 g/dl and no edema?

Ans: 4.783 (higher in edema) *(This is just the slope for the edema covariate)*

g. (5 points) Based on the regression model including both albumin and edema, what is the best estimate for the difference in mean bilirubin between two groups of subjects having the same edema status but who differ in serum albumin by 1.5 g/dl? Provide a confidence interval for this estimate.

Ans: 4.060 (95% CI 0.7165 to 7.404) (higher in low albumin) *(This is just the 1.5 times the slope for the albumin covariate and its CI)*

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

Ans: The intercept is the estimated mean bilirubin in a group of patients with albumin = 0 and not having edema. This is not compatible with life (and that low albumin would certainly have edema), so there is no scientific relevance.

i. (5 points) Provide an interpretation for the slope for the albumin predictor in the regression model including both albumin and edema. What scientific use would you make of this estimate?

Ans: The albumin parameter estimate is the estimated average difference in mean bilirubin per 1 g/dl difference in albumin when comparing two groups that differ in their albumin levels but have similar edema status. This can be used as a measure of the association between bilirubin and albumin after adjustment for edema.

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

Ans: Yes, the p value is 0.018, so with 95% confidence I can state that when comparing patients of similar edema status, there is tendency for bilirubin to be higher in subjects having lower albumin.

k. (5 points) Provide an interpretation for the slope for the edema predictor in the regression model including both albumin and edema. What scientific use would you make of this estimate?

Ans: The edema parameter estimate is the estimated average difference in mean bilirubin when comparing subjects with and without edema but having similar albumin levels. This can be used as a measure of the association between bilirubin and edema after adjustment for albumin.

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

Ans: Yes, the p value is < 0.0005, so with 95% confidence I can state that when comparing patients of similar albumin levels, there is tendency for bilirubin to be higher in subjects having edema than in those without edema..

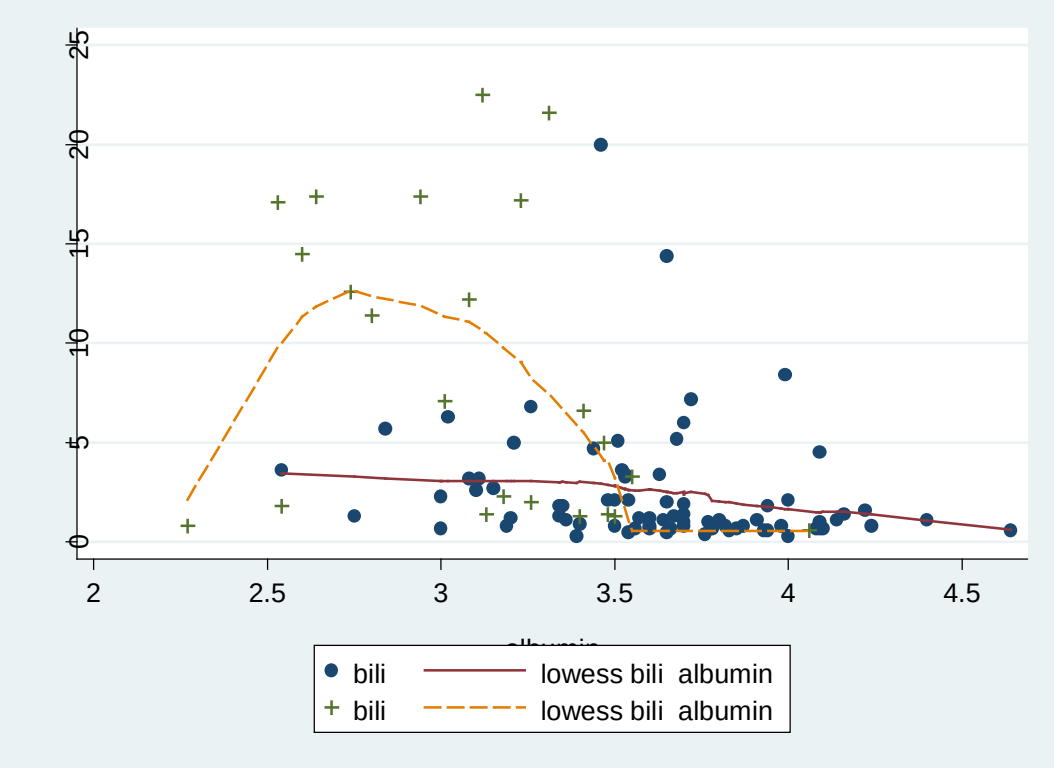
m. (5 points) Based on the regression model including both albumin and edema, what is the best estimate for the average standard deviation of bilirubin measurements within a group that is homogeneous with respect to albumin level and the presence of edema?

Ans: Using the root MSE, we estimate average SD of 4.374 mg/dl in each group defined by albumin and edema status.

n. (5 points) Is there evidence that presence of edema would confound an analysis that merely considered the association between bilirubin and albumin? What would you have to consider?

Ans: There is a tendency toward higher albumin in patients without edema than without, so there does appear to be an association between albumin and edema in the sample. Also, edema is highly predictive of bilirubin after adjusting for albumin. It is therefore not surprising that the magnitude of the association between albumin and bilirubin is markedly different when adjusting or not adjusting for edema. This is diagnostic of confounding in linear regression models so long as edema is not considered in the causal pathway between bilirubin and albumin. *(In primary biliary cirrhosis, the likely causal pathway is that PBC causes elevated bilirubin, and severe PBC leads to decreased albumin, which in turn can cause edema.)*

2. The following scatterplot displays bilirubin (y axis) versus albumin (x axis) within strata defined by no edema (solid points and solid lowess curve) and presence of edema (points marked by + and dashed lowess curve).



a. (5 points) From this plot, comment on the reliability of your answers to parts a through d of problem 1.

Ans: The lowess curve shows marked curvilinearity for the edema stratum. To the extent that a curvilinear relationship would exist in either stratum, we should not feel comfortable using the model which borrows information across groups to predict individual group means. *(If we had used a model that included albumin, edema, and the albumin-edema interaction, then there would have been no borrowing of information across the strata defined by edema. Also: For what it is worth, I don't know how much I really believe there is a curvilinear effect in bilirubin vs albumin in the edema group—this could be an instance of the lowess smooth curving down to meet the lowest observation.)*

b. (5 points) From this plot, comment on the reliability of your answers to parts f, g, h, i, and k of problem 1.

Ans: To the extent that we are interested in estimate a first order trend (so an average difference in mean bilirubin across groups differing in their covariate values), these estimates could still be regarded fairly representative of those first order trend. These estimates might not be exactly correct for all such comparisons, however, due to the curvilinearity in the data.

c. (5 points) From this plot, comment on the reliability of your answers to parts j and l of problem 1.

Ans: The plot shows marked heteroscedasticity (unequal variance) of the bilirubin levels across the different albumin levels. This would make inference based on classical linear regression problematic.

d. (5 points) In problem 1, I also presented a regression model of bilirubin regressed on albumin alone. How might this analysis have changed if I had used robust standard errors?

Ans: The distribution of albumin appears skewed to the left, and the variability of bilirubin appears highest in the groups with low albumin where we have the least data. This would tend to make inference based on classical linear regression anti-conservative. Use of robust standard errors would correct this anti-conservatism. Thus in such an analysis, the parameter estimates would not change, but the confidence intervals would likely be wider and the P values higher. *(When I do this analysis, the 95% CI for the albumin parameter does in fact get wider and is from -7.31 to -2.54.)*

3. Now suppose we consider a log transformation of bilirubin: ***logbili= log(bili)***. Consider the following linear regression analysis.

```
. regress logbili albumin edema in 1/100

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  2,    97) =   20.65
       Model |  34.9186614     2  17.4593307            Prob > F      =  0.0000
    Residual |  81.9972366    97  .845332336           R-squared     =  0.2987
-------------+------------------------------           Adj R-squared =  0.2842
       Total |  116.915898    99  1.18096867           Root MSE      =  .91942

------------------------------------------------------------------------------
     logbili |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     albumin |  -.8377124   .2360884    -3.55   0.001    -1.306283   -.3691421
       edema |   .7307378   .2512203     2.91   0.004     .2321349    1.229341
       _cons |    3.47105   .8617694     4.03   0.000     1.760676    5.181423
------------------------------------------------------------------------------
```

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

Ans: The exponentiated intercept $e^{3.471}$= 32.1 mg/dl is the estimated geometric mean bilirubin in a group of patients with albumin = 0 and not having edema. This is not compatible with life (and that low albumin would certainly have edema), so there is no scientific relevance. *(I accepted you telling me that the intercept was the estimated mean log bilirubin in that group, but I think it far better to interpret the parameter in the same units the measurement was supplied.)*

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

Ans: The exponentiated albumin parameter estimate $e^{-0.8377}$= 0.433 is the estimated ratio of geometric mean bilirubin per 1 g/dl difference in albumin when comparing two groups that differ in their albumin levels but have similar edema status. This can be used as a measure of the association between bilirubin and albumin after adjustment for edema. *(I accepted you telling me that the slope was the average difference in mean log bilirubin between groups, but I think it far better to interpret the parameter in the same units the measurement was supplied.)*

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

Ans: The exponentiated edema parameter estimate $e^{0.7307}$= 2.077 is the estimated ratio of geometric mean bilirubin when comparing subjects with and without edema but having similar albumin levels. This can be used as a measure of the association between bilirubin and edema after adjustment for albumin. *(I accepted you telling me that the slope was the average difference in mean log bilirubin between groups, but I think it far better to interpret the parameter in the same units the measurement was supplied.)*

4. The following table presents the cross classification of a sample with respect to sex, prior cardiovascular disease, and death within 4 years (no subjects are censored).

	Females		**Males**		**All Subjects**	
	Alive	**Dead**	**Alive**	**Dead**	**Alive**	**Dead**
No CVD	2244	116	1317	174	3561	290
CVD	464	80	480	125	944	205
Total	**2708**	**196**	**1797**	**299**	**4505**	**495**

a. (10 points) Suppose we fit a logistic regression modeling the indicator of death within 4 years (response variable) as a function of a variable indicating male sex (so *male*=0 for females and *male*=1 for males). Can you find the intercept and slope for such a model? If so, do so. If not, explain the difficulty.

Ans: When fitting a regression model with a single binary predictor, there is no need to borrow information across groups: The regression parameter estimates will correspond to the relevant sample descriptive statistics. With logistic regression, we model the log odds, so the intercept will be the log of the odds of death for females: log (196 / 2708) = -2.626.

Similarly, the slope will be the log of the odds ratio comparing the odds of death among males to that for females: log ((299 / 1797) / (196 / 2708)) = 0.8324.

b. (10 points) Suppose we fit a logistic regression modeling the indicator of death within 4 years (response variable) as a function of a variable indicating male sex (so male=0 for females and male=1 for males) and an indicator of prior cardiovascular disease (so *cvd*=0 if none, *cvd=1* if so). Can you find the intercept and slopes for both the *male* and *cvd* variables for such a model? If so, do so. If not, explain the difficulty.

Ans: When fitting a regression model with two predictors and without an interaction, we are necessarily borrowing information across groups For instance, the parameter for *male* is trying to estimate the log odds ratio comparing males to females while holding *cvd* constant. But in the data, there will not be the exact same odds ratio in each *cvd* stratum, so some sort of average value is used. And borrowing data in this way will mean that the intercept estimate is also affected by the borrowing of data. The correct estimates are found in a computerized iterative search—there is no simple formula. *(In this case, the logistic regression estimates actually obtained are an intercept of -2.851, a male slope parameter estimate of 0.7406, and a cvd slope parameter estimate of 0.8905. Compare these estimates with the estimates derived for a model with the interaction. By the way: This was the most difficult problem on the exam.)*

c. (20 points) Suppose we fit a logistic regression modeling the indicator of death within 4 years (response variable) as a function of a variable indicating male sex (so male=0 for females and male=1 for males), an indicator of prior cardiovascular disease (so *cvd*=0 if none, *cvd=1* if so) and their interaction *mcvd= male * cvd*.. Can you find the intercept and slopes for both the *male*, *cvd*, and *mcvd* variables for such a model? If so, do so. If not, explain the difficulty.

Ans: When fitting a regression model with two binary predictors and their interaction, there is no need to borrow information across groups: The four parameters (intercept and three slope parameters) can predict the group summary measures perfectly. We can derive the estimates by considering the interpretation of each of the parameters. With logistic regression, we model the log odds, so:

$$\log(odds\, death \mid male, cvd) = \beta_0 + \beta_m \times male + \beta_c \times cvd + \beta_{mc} \times mcvd$$

- **The estimate of the intercept β_0 is the log of the odds of death for females without prior cardiovascular disease: log (116 / 2244) = -2.962.** *(The intercept corresponds to the group with male==0 and cvd==0.)*
- **The estimate of the *cvd* slope β_c is the log of the odds ratio comparing the odds of death among females with prior CVD to that for females without prior CVD (this could also be the difference in the log odds): log ((80 / 464) / (116 / 2244)) = 1.2046.** *(The cvd slope is interpretable as the log odds ratio comparing subjects with CVD to subjects without CVD while holding all other variables constant. The only way we can hold the interaction term constant is if male==0, so we must be considering females.)*
- **The estimate of the *male* slope β_m is the log of the odds ratio comparing the odds of death among males without prior CVD to that for females without prior CVD: log ((174 / 1317) / (116 / 2244)) = 0.9384.** *(The male slope is interpretable as the log odds*

ratio comparing males to females while holding all other variables constant. The only way we can hold the interaction term constant is if cvd==0, so we must be considering subjects without prior CVD.)

- **The estimate of the *mcvd* slope β_{mc} is the log of the ratio of the odds ratio (OR) comparing CVD to non-CVD for males to the OR comparing CVD to non-CVD for females: log ([(125 / 480) / (174 / 1317)] / [(80 / 464) / (116 / 2244)]) = - 0.5260.** *(The interpretation of this parameter cannot use the "holding all other covariates constant" approach, because there is no way to hold both male and cvd constant while changing their product mcvd. The interpretation of an interaction is the "difference of differences" in an additive model and the "ratio of ratios" in a multiplicative model.)*

5. Suppose we are interested in the association between age and death in a population of patients recently admitted to a hospital for cardiovascular disease. Available data include

   - *age:* age of patient in years
   - *prevhosp:* an indicator that the patient had been previously hospitalized for cardiovascular disease (0= no, 1=yes)
   - *obs:* time in years that a patient was followed
   - *death:* an indicator that the patient was observed to die (0= patient was still alive at the time indicated by *obs*, 1= patient was observed to die at the time indicated by *obs*)

   The following proportional hazards analyses were performed on these data:

```
. tabstat age, by(prevhosp) stat(n mean sd min q max) col(stat)
Summary for variables: age
     by categories of: prevhosp
```

prevhosp	N	mean	sd	min	p25	p50	p75	max
0	50	72.56	4.558732	65	69	72	75	85
1	50	72.2	5.43233	65	68	72	74	89
Total	100	72.38	4.992479	65	68	72	75	89

Proportional hazards regression on age

```
. stcox age, robust
         failure _d:  death
   analysis time _t:  obs
Cox regression -- Breslow method for ties
No. of subjects      =          100                 Number of obs   =       100
No. of failures      =           70
Time at risk         =  123.0409999
                                                    Wald chi2(1)    =      2.76
Log pseudolikelihood =   -271.09993                 Prob > chi2     =    0.0966
```

_t	Haz. Ratio	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	1.048262	.029735	1.66	0.097	.9915725	1.108192

Proportional hazards regression on age and previous hospitalization

```
. stcox age prevhosp, robust
         failure _d:  death
   analysis time _t:  obs
Cox regression -- Breslow method for ties
No. of subjects      =          100                 Number of obs   =       100
No. of failures      =           70
Time at risk         =  123.0409999
                                                    Wald chi2(2)    =     42.28
Log pseudolikelihood =   -252.31456                 Prob > chi2     =    0.0000
------------------------------------------------------------------------------
             |               Robust
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.074049   .0260873     2.94   0.003     1.024117    1.126416
    prevhosp |   5.003388   1.335928     6.03   0.000     2.964759    8.443819
------------------------------------------------------------------------------
```

a. (10 points) Based on the model that includes only age, provide the scientific conclusions you would reach about any association between time to death and age. Include estimates and inference.

Ans: From proportional hazards regression, we estimate a 4.83% higher instantaneous risk of death per 1 year difference in age. This is not statistically significant (P= 0.097, 95% CI 0.84% lower to 10.8% higher), so we cannot reject the null hypothesis of no increased risk of death across age groups.

b. (10 points) Based on the model that includes both age and previous hospitalization, provide the scientific conclusions you would reach about any association between time to death and age. Include estimates and inference.

Ans: From proportional hazards regression adjusting for previous hospitalizations, we estimate a 7.40% higher instantaneous risk of death per 1 year difference in age when comparing groups with similar history of prior hospitalization. This is statistically significant (P= 0.003, 95% CI 2.4% to 12.6% higher), so we can with high confidence reject the null hypothesis of no increased risk of death across age groups.

c. (10 points) How would you explain any difference in your results? Is there evidence that previous hospitalization confounds the analysis of an association between age and time to death?

Ans: From the descriptive statistics we see that the age distribution is nearly identical for the patients with no prior history of hospitalization and those having been previously hospitalized for CVD. Hence, prior hospitalization does not confound our assessment of the association between age and risk of death. On the other hand, we do observe that a history of prior hospitalization is an extremely strong predictor of death (age adjusted HR is 5.00). Hence the higher estimate of the hazard ratio for age is attributable to the greater precision afforded by the adjusted model. *(Note the "de-attenuation" of the HR estimate when adjusting for the precision variable: The adjusted HR is 50% larger than the unadjusted HR. It is further of interest to note that the main effect of adjustment for the precision variable is that "de-attenuation": The width of the CI as measured by the ratio of the upper bound to the lower bound is nearly identical: 1.108192 / 0.9915725 = 1.12 versus 1.126416 / 1.024117 = 1.10.)*

6. The following analysis also added a predictor ***agesqr= age ^ 2***.

```
. stcox age agesqr prevhosp, robust
        failure _d:  death
  analysis time _t:  obs
Cox regression -- Breslow method for ties
No. of subjects      =          100              Number of obs   =       100
No. of failures      =           70
Time at risk         =  123.0409999
                                                  Wald chi2(3)    =     52.77
Log pseudolikelihood =   -251.60793               Prob > chi2     =    0.0000
------------------------------------------------------------------------------
             |               Robust
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    .549638   .2950613    -1.11   0.265     .1919233    1.574077
      agesqr |   1.004503   .0035695     1.26   0.206     .9975314    1.011524
    prevhosp |   5.062474   1.341895     6.12   0.000     3.011183    8.511152
------------------------------------------------------------------------------
```

a. (10 points) Based on the above analysis, is there evidence that the effect of age on the log hazard rate is well approximated by a straight line? Explain your reasoning

Ans: The slope parameter estimate for the *agesqr* term is not statistically different from 0 (P = 0.206), so we have no strong evidence for a departure from a straight line model. (At least not a departure that can be detected by a quadratic model.)

b. (Bonus: 10 points) Using on the above analysis, how would you test for an association between age and survival?

Ans: We would need to simultaneously test that the slopes for *age* and *agesqr* are 0 using a multiple partial chi squared test. *(If I perform that test, we find a P value of 0.0004. This illustrates the difficulty in interpreting the P values for individual slopes when several modeled variables are derived from the same scientific variable. In linear regression, we use the F distribution, but in logistic, Poisson, and proportional hazards regression, it is a chi squared statistic that is used.)*

7. A scientific colleague was examining how the relationship between C-reactive protein (CRP, a marker of inflammation) and age differed across the sexes. I would, of course, ideally wanted output from a linear regression of *crp* (C reactive protein) 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*. He brought to me the following output from two linear regressions of CRP on age. From this output (he could not provide the data) he wanted to know the answer to a number of questions.

Linear regression model for females:

```
. regress crp age if male==0, robust
Linear regression                               Number of obs =    2861
                                                F(  1,  2859) =    8.10
                                                Prob > F      =  0.0045
                                                R-squared     =  0.0027
                                                Root MSE      =  5.4835
------------------------------------------------------------------------------
             |               Robust
         crp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0516989    .018163    -2.85   0.004    -.0873128   -.0160849
       _cons |   7.382912   1.346132     5.48   0.000     4.743424    10.0224
------------------------------------------------------------------------------
```

Linear regression model for males:

```
Linear regression                               Number of obs =    2072
                                                F(  1,  2070) =    1.18
                                                Prob > F      =  0.2770
                                                R-squared     =  0.0009
                                                Root MSE      =  6.9576
------------------------------------------------------------------------------
             |               Robust
         crp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0375293    .034512     1.09   0.277    -.0301526    .1052111
       _cons |  -1.354464   2.504879    -0.54   0.589    -6.266809    3.557881
------------------------------------------------------------------------------
```

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

Ans: The intercept in the model with an interaction would correspond exactly to the intercept in the model fit to females only: 7.383.

b. (5 points) Supposing the researcher 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 model with an interaction would correspond exactly to the slope for *age* in the model fit to females only: -0.05170.

c. (5 points) Supposing the researcher 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 model with an interaction would correspond exactly to the difference in the intercepts in the model fit to males only and the model fit to females only: - 1.354 - 7.383 = - 8.737.

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

Ans: The slope for *maleage* in the model with an interaction would correspond exactly to the difference in the slopes for *age* in the model fit to males only and the model fit to females only: 0.03753 – (- 0.05170) = 0.08923.

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

Ans: The standard error for the estimate of the difference in slopes can be computed from the standard errors derived from the two independent samples: We take the square root of the sums of the two squared standard errors: sqrt ($0.034512^2 + 0.018163^2$) = 0.03900. We can then create a Z test: Z = 0.08923 / 0.03900 = 2.2879. Under the null hypothesis, this Z statistic should have a standard normal distribution, so we can compare it to the critical value 1.96 and state with 95% confidence that there is a statistically significant difference between the age slopes for the two sexes.

f. (10 points) Suppose we had really wanted to know the association between CRP and age in the entire population, irrespective of sex. How might you approximate the slope of the age covariate if we had fit a regression model only including age to a sample that was 50% male and 50% female? Would that parameter likely indicate a statistically significant association between CRP and are in the population?

Ans: One approach would be to take the average of the age slopes for the two sexes: So we would estimate an age slope of (-0.0516989 + 0.0375293) / 2 = - 0.0070848.

The standard error for this estimate would be the square root of one-fourth of the sums of the two squared standard errors divided by 4 (recall that when multiplying a statistic by a constant, we increase its variance by a factor of the constant squared): So we would estimate a SE of sqrt (($0.034512^2 + 0.018163^2$) / 4) = 0.01950. We can then create a Z test by dividing the estimated slope by its SE: Z = - 0.0070848 / 0.01950 = - 0.363. Under the null hypothesis, this Z statistic should have a standard normal distribution, so we can compare it to the critical value 1.96 and state with 95% confidence that there is insufficient evidence to reject a null hypothesis of no difference in average CRP across age groups after adjusting for age.
(This was essentially a stratified analysis.)

Grade distribution:

Maximum possible:		235
Highest achieved:		186
Mean (SD):		134 (26)
Percentiles:	90^{th}	172
	80^{th}	154
	50^{th}	135
	20^{th}	109
	10^{th}	104