Bios 312: Modern Regression Analysis

Midterm Examination

February 28, 2013

Name: ___KEY_____

Instructions: Please provide concise answers to all questions. Questions are of varying levels of difficulty, 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 one (1) page 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 are running short of time, leave answers in unsimplified form to receive the majority of credit.

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

Signature: _____

This exam consists of **11** pages including an Appendix of Results. There are 140 total points.

- Question 1: Parts 1-7, 60 points
- Question 2: Parts 1-4, 40 points
- Question 3: Parts 1-4, 40 points

Stem and Leaf Plot of Grade Distribution

8	| 0 4	Mean: 105
9	| 1 2	Median: 106
10	| 3 5 6 6 7	
11	| 2 4 6	
12	| 1	
13	| 7	

Question 1 (60 points) Use the results in the Appendix to answer the following questions about the association of serum arginine (arg) with cholesterol (chol), age, and male sex. Assume the assumption of homoscedasticity holds for all models.

1-1. Based on the simple linear regression model (only including cholesterol as covariate), what scientific conclusions can you make about the association between cholesterol and arginine? Provide an interpretation for the cholesterol coefficient. Include estimates of the effect size and statistical inference in your written description of the results. (10 points)

Comparing two subjects who differ in their cholesterol level by one unit, a 1-unit increase in cholesterol is associated with a 3.46 unit change in serum arginine. We are 95% confident that the true change in arginine per unit change in cholesterol is between 1.22 and 5.68. Because the confidence interval does not contain 0 (and p = 0.003), there is a significant association between cholesterol and arginine at the 5% significance level.

1-2. Based on the multivariable linear regression model with cholesterol, age and male as covariates, what scientific conclusions can you make about the association between cholesterol and arginine? Provide an interpretation for the cholesterol coefficient. Include estimates of the effect size and statistical inference in your written description of the results. (10 points)

Comparing two subjects who differ in their cholesterol level by one unit but have the same age and gender, a 1-unit increase in cholesterol is associated with a 2.35 unit increase in serum arginine. We are 95% confident that the true change in arginine per unit change in cholesterol is between 0.36 and 4.35. Because the confidence interval does not contain 0 (and p = 0.02), there is a significant association between cholesterol and arginine at the 5% significance level.

1-3. Is there any evidence that age confounds the association between arginine and cholesterol? Do you consider age to be a precision variable? Explain your answers. (10 points)

A confounder must be related to both the outcome (causally) and the predictor of interest. Symptoms of confounding in a linear regression model are scientifically meaningful (e.g. 10%) changes in the regression coefficient estimates. When we compare the coefficient for cholesterol in the unadjusted model to the model where we adjust for age, the coefficient changes from 3.46 to 2.40; this large change is indicative of confounding. To evaluate a precision variable, we should look at the standard error of the cholesterol coefficient. It changes from 1.12 (unadjusted) to 1.28 (adjusted); because it increases, there is little evidence that it is a precision variable.

1-4. Is there any evidence that male sex confounds the association between arginine and cholesterol? Do you consider male sex to be a precision variable? Explain your answers. (10 points)

Using the same criteria as above, male sex appears to be a precision variable. When we adjust for male sex, the standard error of the cholesterol coefficient decreases from 1.12 (unadjusted) to 0.89 (adjusted for male); also note the decrease in the RMSE from 14.5 to 11.5. The estimate of the regression coefficient changes very little (3.46 vs 3.44), so there are not strong symptoms of confounding. Note that this is not definitive proof of lack of confounding as confounding is primarily a concept that does not have a formal statistical test.

1-5. Based on the multivariable linear regression model with cholesterol, age and male as covariates, what is the expected value of arginine when cholesterol is 185, sex is male, and age is 50? What assumption about the form of the relationship between arginine and cholesterol is needed for this to be interpreted as either an expected value or a predicted value of a new observation? (10 points)

-422.97 + 185*2.356 + 50*0.3883 +17.82 = 50.1

We assume that the relationship between arginine and cholesterol follows a straight line.

1-6. Briefly describe what the assumption of homoscedasticity means for the multivariable linear regression model with cholesterol, age and male as covariates. (5 points)

Homoscedasticity means that the residuals have constant variance for all expected values of Y. Since the expected value is a function of age, cholesterol and male, the variance will be the same across groups defined by these three covariates.

1-7. Suppose you fit the simple linear regression model (only age as a covariate) and use robust standard error estimates. How would this alter your model results? In particular, discuss changes to your estimates of the slope and the standard error of the slope. You are not given this output, so just discuss the changes you expect to occur. (5 points)

The regression coefficient estimates would not change. The standard error estimates would increase because we are told to assume that the assumption of homoscedasticity holds. When this assumption holds, the classical model will be more efficient than using robust standard errors.

Question 2 (40 points): Suppose that we are interested in examining the association of time to death with age and smoking status in a group of elderly patients enrolled in a cohort study. Use the analysis conducted in the appendix to answer the following questions.

2-1) Based on the logistic regression model, what scientific conclusions can you make about the association between age and death? If appropriate, include estimates of the effect size and statistical inference in your written description of the results. (10 points)

No scientifically relevant conclusions can be drawn. This is censored survival data, so logistic regression is not appropriate because it does not take into account the variable time under observation.

(Aside: In this dataset-- although you would have no way of knowing this-- subjects with the earliest censoring times were minorities because a second cohort of individuals was recruited to enhance minority representation. In this case, failure to adjust for the censoring could lead to very different estimates of survival probabilities if age and/or its association with survival differed across race/ethnicity. There are many nuances here, but the general point remains: with censored observations, you need to consider censoring to draw valid scientific conclusions).

2-2) Based on the simple PH regression model (only age as a covariate), what scientific conclusions can you make about the association between age and time to death? If appropriate, include estimates of the effect size and statistical inference in your written description of the results. (10 points)

Every one unit increase in age is associated wit a 1.03 fold increase in the instantaneous risk of death. We are 95% confident that the true hazard ratio is between 1.01 and 1.06. Since the confidence interval does not contain 1, and p = 0.014, we can conclude that there is a significant association between age and the instantaneous risk of death.

2-3) Based on the multivariable PH regression model (age and smoker as covariates), what scientific conclusions can you make about the association between age and time to death? If appropriate, include estimates of the effect size and statistical inference in your written description of the results. (10 points)

Among subjects with the same smoking status but differing in age, the subject who is 1 year older is at a .04 fold increased risk (hazard) of death compared to the younger subject. We are 95% confident that the true adjusted hazard ratio is between 1.01 and 1.07. Since the confidence interval does not contain 1 (and p = 0.008), we conclude that there is a significant association between age and the risk of death, holding smoking status constant.

2-4) Compare the simple and multivariable PH regression models. Is there evidence that smoking status confounds the association between age and time to death? Is there evidence that it is a precision variable? Justify your answer. (10 points)

To evaluate symptoms of confounding, we evaluate if the hazard ratio estimate changes by a scientifically important amount from the unadjusted to the adjusted model. Here, it changes from 1.034 to 1.037, so there is little evidence of confounding. When modeling hazards (or probabilities), precision variables will tend to drive the coefficient estimate away from the null. In this case, we see the hazard ratio increase so there is some evidence that smoking status is a precision variable. Furthermore, smoking status is significantly associated with the risk of death (p =0.020), which is also consistent with smoking being a precision variable.

Note that I am primarily concerned with your justification of your answer, which should raise the issues stated above. If, for example, you followed the same logic and concluded that smoker was not a precision variable, I would consider that acceptable in this case. In real observational data, it is rare to find that a variable acts purely as either a precision variable or a confounder.

Question 3 (40 points): The following table presents the cross-classification table of sex, diabetic status, and presence of atherosclerotic disease.

	Females		**Males**		**All Subjects**	
	No Disease	**Disease**	**No Disease**	**Disease**	**No Disease**	**Disease**
Not Diabetic	70	24	52	19	122	43
Diabetic	12	8	12	11	24	19
Total	**82**	**32**	**64**	**30**	**146**	**62**

We will consider three models for p, the probability of atherosclerotic disease

- Model 3.A: $\log(p/(1-p)) = \alpha_0 + \alpha_1$*male
- Model 3.B: $\log(p/(1-p)) = \beta_0 + \beta_1$*male + β_2*diabetic
- Model 3.C: $\log(p/(1-p)) = \gamma_0 + \gamma_1$*male + γ_2*diabetic + γ_3*diabetic*male

3-1) Before considering the models, calculate the following conditional probabilities using the data in the table. Report as either fractions or decimals. Note that the '|' symbol should be read as 'given'; e.g. Prob(Disease | Male) is "The probability of disease given male". (10 points)

Prob(Disease | Male) = ***30/(30+64) = 0.3191***

Prob(Disease | Female) = ***32/(32+82) = 0.2807***

Prob(Disease | Male and Diabetic) = ***11/(11+12) = 0.4782***

Prob(Disease | Female and Diabetic) = ***8/(8+12) = 0.4000***

3-2) Suppose we want to fit a simple logistic regression model (Model 3.A) with atherosclerotic disease (0=No Disease, 1=Disease) as the outcome and an indicator for male sex (0=Female, 1=Male) as the predictor. Using the above table, can you find the intercept and slope for such a logistic regression model? If so, what are the estimates of the intercept and slope? If not, explain the difficulty. (10 points)

α_0 is the log odds of disease in females. Using the results from part 1, the odds of disease in females is 0.2807/(1-0.2807) = 0.3902; the log odds is -0.941. This is α_0

$\alpha_0 + \alpha_1$ is the log odds of disease in males. Using the results from part 1, odds in males =0.3191 / (1-0.3191) = 0.4686; the log odds is -0.7579. This is $\alpha_0 + \alpha_1$, α_1 is thus 0.183.

Another solution, among many, would be to realize that exp(α_1) is the odds ratio comparing males to female, so α_1 is the log odds ratio. log(82*30/(32*64)) = 0.183

3-3) Suppose we want to fit a multiple logistic regression model (Model 3.B) with atherosclerotic disease as the outcome, and indicators for male sex and diabetic status (0=Not Diabetic; 1 = Diabetic) as the adjustment variables. Using the above table, can you find the intercept and slopes for such a logistic regression model? If so, what are the estimates of the intercept and slopes? If not, explain the difficulty. (10 points)

These estimates cannot be obtained from the given data. In this model β_1 is the change in the log odds of disease comparing a male to a female subject while holding diabetic status constant. Thus, β_1 is a approximately a weighted average of male effect in non-diabetics and the male effect in diabetics. Without knowing the precise weight used by logistic regression, the estimate β_1 of cannot be computed. Similarly β_2 is the change in the log odds comparing diabetics to non-diabetics, averaging over smoking status; without knowing the weights, we cannot compute β_2.

3-4) Suppose we want to fit a multiple logistic regression model (Model 3.C) with atherosclerotic disease as the outcome, and indicators for male sex, diabetic status, and the interaction of male sex and diabetic status as the adjustment variables. Using the above table, can you find the intercept and slopes for such a logistic regression model? If so, what are the estimates of the intercept and slopes? If not, explain the difficulty. (10 points)

Yes, they can be determined. Parameters in the model have the following interpretations:

γ_0 is the log odds of disease in female non-diabetics
= log(0.2553/1-.2553) = -1.071
$\gamma_0 + \gamma_1$ is the log odds of disease in male non-diabetics
= log(0.2676/1-.2676) = -1.007
$\gamma_0 + \gamma_2$ is the log odds of disease in female diabetics
= log(0.4/1-.4) = -0.4054
$\gamma_0 + \gamma_1 + \gamma_2 + \gamma_3$ is the log odds of disease in male diabetics
= log(0.4782/1-.4782) = -0.0872

The above calculations use the conditional probabilities in part 1, plus two additional probabilities calculated in a similar manner. Using these results, we find that

γ_0 = -1.07
γ_1 = -1.007 - (-1.071) = 0.064
γ_2 = -0.4054 - (-1.071) = 0.665
γ_3 = -0.0872 - (-1.071 + 0.064 + 0.6656) = 0.255

End of Exam Questions.

Appendix for Question 1

```
. tabstat chol age male, stat(n mean sd min p25 p50 p75 max) format(%9.2f)

   stats |      chol       age      male
---------+------------------------------
       N |     75.00     75.00     75.00
    mean |    185.39     53.41      0.45
      sd |      1.51      8.30      0.50
     min |    182.03     40.28      0.00
     p25 |    184.11     46.97      0.00
     p50 |    185.48     54.55      0.00
     p75 |    186.46     60.31      1.00
     max |    189.17     69.76      1.00
----------------------------------------



. regress arg chol

      Source |       SS       df       MS              Number of obs =      75
-------------+------------------------------           F(  1,    73) =    9.55
       Model |  2011.99875     1  2011.99875           Prob > F      =  0.0028
    Residual |  15383.8822    73  210.738112           R-squared     =  0.1157
-------------+------------------------------           Adj R-squared =  0.1035
       Total |  17395.8809    74  235.079472           Root MSE      =  14.517

------------------------------------------------------------------------------
         arg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        chol |   3.455854   1.118441     3.09   0.003     1.226804    5.684904
       _cons |  -598.0249   207.3599    -2.88   0.005    -1011.293   -184.7571
------------------------------------------------------------------------------



. regress arg chol age

      Source |       SS       df       MS              Number of obs =      75
-------------+------------------------------           F(  2,    72) =    6.22
       Model |  2563.65671     2  1281.82835           Prob > F      =  0.0032
    Residual |  14832.2242    72  206.003114           R-squared     =  0.1474
-------------+------------------------------           Adj R-squared =  0.1237
       Total |  17395.8809    74  235.079472           Root MSE      =  14.353

------------------------------------------------------------------------------
         arg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        chol |   2.396671   1.281303     1.87   0.065    -.1575602    4.950903
         age |   .3809583   .2327981     1.64   0.106    -.0831162    .8450328
       _cons |  -422.0045   231.5209    -1.82   0.072     -883.533    39.52406
------------------------------------------------------------------------------
```

Appendix for Question 1 (cont.)

```
. regress arg chol age male

      Source |       SS       df       MS              Number of obs =      75
-------------+------------------------------           F(  3,    71) =   22.44
       Model |  8465.65039     3  2821.88346           Prob > F      =  0.0000
    Residual |  8930.23055    71  125.777895           R-squared     =  0.4866
-------------+------------------------------           Adj R-squared =  0.4650
       Total |  17395.8809    74  235.079472           Root MSE      =  11.215

------------------------------------------------------------------------------
         arg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        chol |   2.356197   1.001209     2.35   0.021     .3598431    4.352552
         age |   .3882987   .1819082     2.13   0.036     .0255841    .7510133
        male |   17.82004   2.601426     6.85   0.000     12.63294    23.00713
       _cons |  -422.9713   180.9071    -2.34   0.022    -783.6898   -62.25271
------------------------------------------------------------------------------



. regress arg chol male

      Source |       SS       df       MS              Number of obs =      75
-------------+------------------------------           F(  2,    72) =   29.90
       Model |  7892.54863     2  3946.27432           Prob > F      =  0.0000
    Residual |  9503.33231    72  131.990726           R-squared     =  0.4537
-------------+------------------------------           Adj R-squared =  0.4385
       Total |  17395.8809    74  235.079472           Root MSE      =  11.489

------------------------------------------------------------------------------
         arg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        chol |   3.435826   .8851478     3.88   0.000     1.671316    5.200336
        male |   17.78733   2.664854     6.67   0.000     12.47504    23.09962
       _cons |  -602.3752   164.1075    -3.67   0.000    -929.5176   -275.2329
------------------------------------------------------------------------------
```

Appendix for Question 2

The analysis of time to death with age and smoking status consider the following variables:

- *death*: Indicator of death status (1=died on study; 0=alive at end of study)
- *ttodth*: Number of days from enrollment to either death or the end of the study
- *age*: Age in years
- *smoker*: Current smoking status (1=Current smoker; 0=Non-current smoker)

```
variable   | N    mean  sd    min p25  p50  p75  max
-----------+--------------------------------------------
death      | 196  0.24  0.42  0   0    0    0    1
ttodth     | 196  2370  702   86  2074 2699 2826 2941
age        | 196  72.9  5.16  65  69   72   77   88
smoker     | 196  0.11  0.32  0   0    0    0    1
-------------------------------------------------------
```

The following three regression models (some output omitted) were run:

Model 1: Logistic regression adjusting for age and smoking status

```
. logistic death age smoker, robust

------------------------------------------------------------
       |             Robust
death  | Odds Ratio Std. Err. z    P>|z|  [95% Conf. Int.]
-------+----------------------------------------------------
age    | 1.203      0.04512   4.95 0.000  1.118  1.295
smoker | 4.142      2.10962   2.79 0.005  1.526  11.23
------------------------------------------------------------
```

Appendix for Question 2 (cont.)

Model 2: Proportional Hazards Regression adjusting for age

```
. stset ttodth death
. stcox age, robust

------------------------------------------------------------------------------
       |  Haz.   Robust
_t     |  Ratio  Std. Err.      z    P>|z|     [95% Conf. Interval]
-------+----------------------------------------------------------------------
age    |  1.034   0.0141     2.46   0.014     1.007    1.062
------------------------------------------------------------------------------
```

Model 3: Proportional Hazards Regression adjusting for age and smoking status

```
. stcox age smoker, robust

------------------------------------------------------------------------------
       |  Haz.   Robust
_t     |  Ratio  Std. Err.      z    P>|z|     [95% Conf. Interval]
-------+----------------------------------------------------------------------
age    |  1.037   0.0144     2.63   0.008     1.009    1.066
smoker |  1.841   0.4836     2.32   0.020     1.100    3.081
------------------------------------------------------------------------------
```