Biost 518
Emerson, Winter 02
February 8, 2002

Biost 518
Applied Biostatistics II

Midterm Examination Key

Name: _____

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.

Appendix A contains data analysis results from a large cohort study involving 735 subjects aged 65 and older. Data available at study entry included demographic data, as well as selected measures of known risk factors for cardiovascular and cerebrovascular disease. Each subject also had an MRI exam of their brain which was measured for the evidence of cerebral atrophy (shrinking of the brain) on a scale from 0 (none) to 100 (severe atrophy). All subjects were also followed a minimum of 5 years for survival. Variables available for this analysis includes:

- mridate= The date on which the participant underwent MRI scan in MMDDYY format.
- age= Participant age at time of MRI (years)
- male= Indicator of whether participant is male (0= female,1=male)
- race= Indicator of participant's race (1= white, 2= black, 3= Asian, 4= other)
- diabetes= Indicator of whether the participant had been diagnosed with diabetes prior to MRI (0= no, 1= borderline, 2= full diabetes). Diabetes is a disease in which a patient does not regulate his/her blood glucose in a normal fashion. Glucose is the main energy source for our bodies, and in diabetes, the cells lose the ability to take glucose from the blood. Persons with diabetes are at high risk of blindness, kidney disease, heart disease, and other diseases of the circulation system.
- genhlth= An indicator of the participant's view of his/her own health (1= excellent, 2= very good, 3= good, 4= fair, 5= poor).
- atrophy= A measure of global brain atrophy detected on MRI. A 0 means little or no shrinkage of the brain, and 100 is marked atrophy of the cerebrum.
- obstime= The total time (in days) that the participant was observed on study between the date of MRI and death or September 16, 1997, whichever came first.
- death= An indicator that the participant was observed to die while on study. If death=1, the number of days recorded in obstime is the number of days between that participant's MRI and his/her death. If death=0, the number of days recorded in obstime is the number of days between that participant's MRI and September 16, 1997.

1. 20 points In Appendix A, cross out any descriptive statistics which do not provide scientifically meaningful information about the dataset. In the space below briefly explain the main factors that influenced

your decisions.

Ans: None of the numerical descriptive statistics are appropriate for *mridate* (in the MMDDYY format, it is best thought of as a nominal variable), for *race* (it is an unordered categorical variable), for *obstime* (it is a right censored measurement), or for *death* (it is measured over varying time frames). The mean and standard deviation are not of much interest for *genhlth* (it is an ordered categorical variable– note that the quantiles are OK here). Only the mean is of much interest for the binary variable of *male* (the mean would be the proportion that is male). Because *diabetes* only had two categories represented, we could consider this a binary variable (otherwise it is an ordered variable). Of the scatterplots, we have no interest in the plot involving *obstime*, because those measurements are censored.

2. 10 points From the information available in Appendix A, is there any evidence that would suggest that classical linear regression would be inappropriate for detecting associations between mean DSST score and degree of cerebral atrophy? Explain your reasoning?

Ans: I saw no marked heteroscedasticity, nor marked curvilinearity. It looks OK to me.

3. 5 points each part Appendix B contains the results of analyses that might be used to explore the association between mean DSST scores and degree of cerebral atrophy. (For this problem, you may assume that necessary assumptions are satisfied, except as noted.)

   a. What is the best estimate for the mean DSST score among participants with an atrophy grade of 20?

Ans: $40.82452 + .0064855 \times 20 = 40.954$**.**

   b. What is the best estimate for the mean DSST score among participants with an atrophy grade of 25?

Ans: $40.82452 + .0064855 \times 25 = 40.987$**.**

   c. What is the best estimate for the difference in mean DSST scores between subjects with atrophy grades of 21 and atrophy grades of 20?

Ans: The estimated slope is .0064855.

   d. What is the best estimate for the difference in mean DSST scores between subjects with atrophy grades of 50 and atrophy grades of 20? Provide a 95% confidence interval for this difference.

Ans: The point estimate is $30 \times .0064855 = 0.1946$**. We obtain the confidence interval by multiplying the confidence interval for the slope by 30: (-1.990, 2.379).**

   e. Provide an interpretation of the intercept from the regression model, including the statistical and scientific relevance of the estimate.

Ans: The regression model estimates that the average DSST score for persons with 0 cerebral atrophy would be 40.82. The 95% confidence interval suggests that this sort of data might reasonably be obtained when the true average DSST was between 38.06 and 43.59. We are highly confident that the true mean is different from 0 (P $<$ .0005). I note that we had a person in our data with an atrophy score as low as 5, so this estimate is not extrapolating very far outside the range of data.

   f. Provide an interpretation of the slope from the regression model, including the statistical and scientific relevance of the estimate.

Ans: The regression model estimates that the difference in average DSST is .00649 between two people differing by 1 in their atrophy scores. Such a result is not atypical of what we might expect when the true difference is 0 (P = 0.861). The 95% confidence interval suggests that this sort of data might reasonably be obtained when the true difference in average DSSTs was between -0.0663 and 0.0793. Thus I would conclude that I do not have evidence to state with high confidence that there is an association

between atrophy and scores on the DSST.

g. Based on the results of the analysis, what would be your conclusion regarding the existence of an association between mean DSST score and atrophy detected on MRI exam? Explain.

Ans: Based on the statistical inference described in part g, I would conclude that I do not have evidence to state with high confidence that there is an association between atrophy and scores on the DSST.

h. Based on the results of the analyses presented in the appendix, what can you say about the presence of a statistically significant correlation between DSST and atrophy grades? Explain.

Ans: The test for a significant linear regression slope is exactly equivalent to a test for a statistically significant correlation. Thus, I would not reject the null hypothesis that the true correlation is 0.

i. Based on the results of the analyses presented in the appendix, what can you say about the presence of a statistically association between mean atrophy grades and DSST score? Explain.

Ans: If I were to perform a regression modeling mean atrophy grades as predicted by DSST scores, the statistical test for a nonzero regression slope would be exactly the same as the test for the correlation described in part h. Thus, I would not reject the null hypothesis that there was no linear trend in mean atrophy across groups defined by DSST scores.

j. Suppose (for this problem only) that the 735 observations in this data actually represented repeat measurements over a 10 year interval on 300 independent subjects. How might this information be expected to affect your answers to the above questions?

Ans: The possible dependence among the individual measurements would mean that the classical linear regression might provide incorrect inference. If people tended toward greater atrophy over time, then the fact that there was variation in the predictor of interest over time and a likely positive correlation among the DSST scores would mean that the results presented in Appendix B was conservative (the reported standard error estimate would be too large). On the other hand, if the atrophy grade for each person was relatively constant over time, the likely positive correlation among the DSST scores would mean that the results presented in Appendix B was anti-conservative (the reported standard error estimate would be too small).

3. Appendix C contains the results of several regression analyses which might be used to explore the association between cerebral atrophy detected on MRI and patient survival. For each of the following analyses, indicate whether the analysis is appropriate to address this question. If it is not, briefly explain why not. If it is, provide a very brief interpretation of the slope (just enough to show me you know what it estimates). In this problem, the variable DeadIn5 was defined as an indicator of death within 5 years of MRI exam.

a. A linear regression of observation time (response) on atrophy (predictor).

Ans: This would be inappropriate because the measurement of observation time was censored.

b. A linear regression of atrophy (response) on an indicator of death within 5 years

Ans: This would be appropriate, because we had at least five years of follow-up on all individuals. This would correspond to the t test which presumes equal variances. The slope estimates the difference in mean atrophy scores between those who died within five years and those who survived at least five years. This model could give inaccurate inference about the difference in means if the variability of atrophy grades was not the same in both groups.

c. A linear regression of atrophy (response) on an indicator of death within 5 years using robust

standard error estimates

Ans: This would be appropriate, because we had at least five years of follow-up on all individuals. This would correspond to the t test which allows unequal variances. The slope estimates the difference in mean atrophy scores between those who died within five years and those who survived at least five years. This model will give accurate inference about the difference in means whether the variability of atrophy grades was the same in both groups or not.

d. A logistic regression of an indicator of death within 5 years (response) on atrophy (predictor)

Ans: This would be appropriate, because we had at least five years of follow-up on all individuals. The exponentiated slope estimates the ratio comparing the odds of death within five years between two groups that differ by one unit in their atrophy score. This model could give inaccurate inference about a trend in the odds of death in 5 years across atrophy groups if the true relationship in the log odds ratio is nonlinear across atrophy groups.

e. A logistic regression of an indicator of death within 5 years (response) on atrophy (predictor) using robust standard error estimates

Ans: This would be appropriate, because we had at least five years of follow-up on all individuals. The exponentiated slope estimates the ratio comparing the odds of death within five years between two groups that differ by one unit in their atrophy score. This model does not give invalid inference even when the true relationship in the log odds ratio is nonlinear across atrophy groups.

f. A proportional hazards regression model of time to death on atrophy (predictor)

Ans: This would of course be entirely appropriate as a method of analysis for these right censored data. The exponentiated slope estimates the ratio comparing the instantaneous risk (hazard) of death between two groups that differ by one unit in their atrophy score. In the absence of using the robust standard error estimates, this model could give inaccurate inference about a trend in the instantaneous risk of death across atrophy groups if the true relationship in the log hazard ratio is nonlinear across atrophy groups or if the proportional hazards assumption does not hold.

4. What would be the conclusions from a t test performed comparing mean atrophy across groups defined by 5 year survival? Explain your answer.

Ans: As noted in the answers to 3b and 3c, those analyses correspond to the t tests which presume equal variance and which allow unequal variance, respectively. I would prefer the t test which allows unequal variances, thus my conclusion is that the observed data are not typical of what would be expected when there is no difference in mean atrophy grade between those who die within 5 years and those who survive at least 5 years ($P < .0005$). We estimate that the average atrophy grade is 5.97 points higher among those who die within 5 years (95% CI 3.43 to 8.50 points higher).

5. Using the results presented in the appendices:

a. What is the best estimate for the probability of 5 year survival among participants with an atrophy grade of 20?

Ans: From the logistic regression, we can estimate the log odds of dying within 5 years for people with atrophy scores of 20 as $-2.814668 + .0345403 \times 20 = -2.123862$**. We then find the odds of dying within 5 years by expeonentiating:** $\exp(-2.123862) = 0.119569$**. The probability of dying within 5 years is then found by computing the odds divided by 1 plus the odds:** $0.119569/(1 + 0.119569) = 0.1068$**. Because the question asked for the probability of survival, we just subtract the probability of dying from 1:** $1 - .1068 = 0.8932$**.**

b. What is the best estimate for the probability of 5 year survival among participants with an atrophy grade of 25?

Ans: By performing the exact same steps using an atrophy grade of 25 we find a 5 years survival probability of 0.8756.