Biost 518: Applied Biostatistics II
Emerson, Winter 2006

Homework #2 Key
January 23, 2006

In this key, I have at times included the Stata code used to solve the problems, along with the Stata output. I have used a blue font to indicate this material that I would not have wanted you to have included this with your homework.

Written problems due at the beginning of class, Friday, January 20, 2006.

Questions 1 - 4 investigate whether the distribution of cerebral atrophy scores differs by age in the population of older patients from which the MRI measurements were sampled. The data is posted on the class web pages.

1. Provide suitable descriptive statistics related to the distribution of atrophy scores by age.

   Answer: Table 1 presents descriptive statistics for atrophy scores within 5 year age intervals. Figure 1 displays a scatterplot of atrophy scores versus age with a superimposed lowess smooth. From the tabulated means, there appears to be a tendency for higher means in the older age groups, with a general trend towards approximately three points higher means for each five years of age (note that I am discounting the estimates in the highest age group which has only two observations). There does not seem to be any particular trend toward increased variability of measurement across the age groups. The general trends seen in the tabulate data are also born out in the scatterplot.

Table 1: Descriptive statistics for atrophy scores by 5 year age interval.

Age	N	Mean	Std Dev	Min	25%ile	Mdn	75%ile	Max	% > 30
65 – 69	117	31.68	11.35	9.00	23.0	31.0	38.0	75.0	51.3
70 – 74	305	34.44	12.70	7.00	25.0	32.0	42.0	84.0	59.0
75 – 79	187	36.93	12.30	14.00	28.0	37.0	46.0	73.0	68.4
80 – 84	81	39.72	11.61	5.00	33.0	38.0	48.0	69.0	80.2
85 – 89	35	45.29	14.45	19.00	36.0	45.0	53.0	81.0	82.9
90 – 94	8	47.00	11.14	35.00	38.0	46.5	51.5	69.0	100.0
95 - 99	2	77.50	4.95	74.00	74.0	77.5	81.0	81.0	100.0
Total	735	35.98	12.92	5.00	27.0	35.0	44.0	84.0	64.2

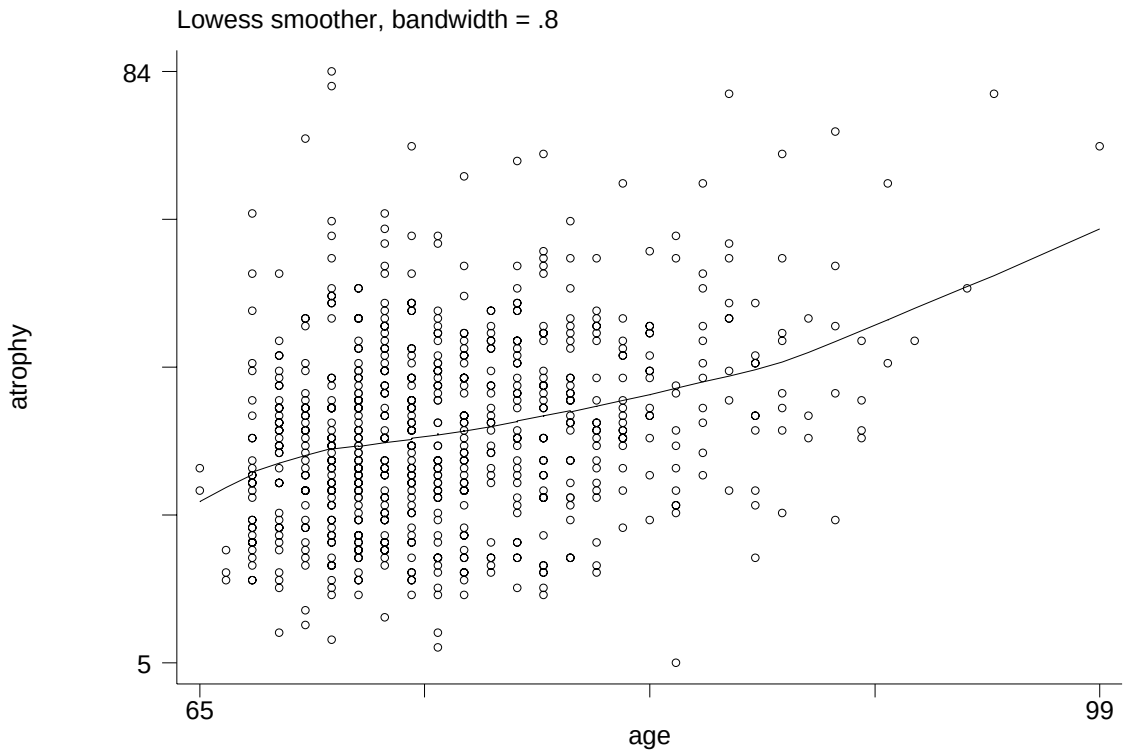

Figure 1: Scatterplot of atrophy scores versus age with superimposed lowess curve.

2. Perform an analysis to determine whether there is a linear trend in mean atrophy scores across age groups.

```
. regress atrophy age, robust
Linear regression                               Number of obs =     735
                                                F(  1,   733) =   60.12
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0867
                                                Root MSE      =  12.359

------------------------------------------------------------------------------
             |               Robust
     atrophy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .6979831   .0900192     7.75   0.000      .521257    .8747093
       _cons |  -16.06213   6.700595    -2.40   0.017    -29.21677   -2.907482
------------------------------------------------------------------------------
```

a. Provide an interpretation of the intercept from the regression model, along with statistical inference pertaining to the hypothesis that the intercept might be 0. What is the scientific relevance of the answer to this question?

<u>Answer</u>: From the regression model, we estimate an intercept of –16.06, which is interpretable as an estimate of the mean atrophy scores in a population of newborn (age 0). The t test based on the estimated intercept and its standard error suggests that we can with high confidence reject the null hypothesis that the mean atrophy scores in newborns might be 0 (P = 0.017). However, given that we had no patients under 65 in our study, we should be extremely reluctant to use this data to try to estimate what the mean atrophy score would be in newborns. Certainly we should be reluctant to believe the inference suggests that with high confidence newborns

would have impossibly low values. So the intercept is not of scientific interest for its own sake.

b. Provide an interpretation of the slope from the regression model, along with statistical inference pertaining to the hypothesis that the slope might be 0. What is the scientific relevance of the answer to this question?

Answer: From the regression model, we estimate a slope of 0.698 which is interpretable as an estimate of the difference in mean atrophy scores between two groups of elderly adults that differ in age by 1 year. The t test based on the estimated slope and its standard error suggests that we can with high confidence reject the null hypothesis that there is no general tendency for the mean atrophy scores to differ in a linear fashion across age groups (P < .0005). Based on the 95% CI, we can with confidence state that the average difference between mean atrophy for two groups differing in age by 1 year is no less than a difference of 0.521 per year difference in age nor more than a difference of 0.875 per year difference in age. That is, the observed estimate of a 0.698 mean difference per year difference in age is not atypical of what we might expect to see if the true mean difference were any number between 0.521 and 0.875.

c. Using the regression model estimated in your analysis, estimate the mean atrophy for 70 year olds and for 90 year olds. Also provide estimates of the standard deviation of atrophy scores within each group. Do you believe that such estimates would accurately reflect the true distribution of atrophy scores in the populations of all 70 or 90 year olds? Why or why not?

Answer: From the regression model, we estimate mean atrophy scores for 70 year olds as –16.06 + 70 * .698 = 32.8, and we estimate the mean atrophy scores for 90 year olds as –16.06 + 90 * .698 = 46.8. There does not seem to be a particularly bad nonlinear trend in the association between atrophy and age, thus these estimated means are probably good approximations. (Note the relatively close agreement with the descriptive statistics given in problem 1.)

We estimate the standard deviation in each group using the root MSE of the residuals: 12.359. This would only be an accurate reflection of the SD within each age group if 1) the means truly appeared to follow a straight line, and 2) the variance in each age group was approximately the same. As noted above, I am not too bothered by any nonlinearity in the association. From the scatterplot, there does seem to be a little less variability in the oldest subjects, but we do not have a very large sample size. From the descriptive statistics in table 1, the SD across age groups is relatively constant. (Note that we would have had a hard time judging this if we had not used age groups based on a fixed number of years.)

3. Perform an analysis to determine whether there is a trend in the geometric mean of cerebral atrophy scores across age groups.

```
. g logatrophy= log(atrophy)
. regress logatrophy age, robust
Linear regression                               Number of obs =     735
                                                F(  1,   733) =   64.81
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0731
```

```
                                                      Root MSE      =  .37367
------------------------------------------------------------------------------
             |               Robust
  logatrophy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0192338   .0023891     8.05   0.000     .0145435    .0239242
       _cons |   2.079393   .1795708    11.58   0.000     1.726859    2.431927
------------------------------------------------------------------------------
```

a. Provide an interpretation of the intercept from the regression model, along with statistical inference pertaining to the hypothesis that the intercept of the regression model might be 0. What is the scientific relevance of the answer to this question?

Answer: From the regression model, we estimate an intercept of 2.079, from which we estimate the geometric mean atrophy scores in a population of newborns (age 0) as 8.00. The t test based on the estimated intercept and its standard error suggests that we can with high confidence reject the null hypothesis that the geometric mean atrophy scores in newborns might be 1.00 (P < 0.0005). However, given that we had no patients under 65 in our study, we should be extremely reluctant to use this data to try to estimate what the geometric mean atrophy score would be in newborns. So the intercept is not of scientific interest for its own sake.

b. Provide an interpretation of the slope from the regression model, along with statistical inference pertaining to the hypothesis that the slope might be 0. What is the scientific relevance of the answer to this question?

Answer: From the regression model, we estimate a slope of 0.0192 from which we estimate the ratio of geometric mean atrophy scores between two groups of elderly adults that differ in age by 1 year as 1.019 (or a 1.9% higher geometric mean for each year difference in age). The t test based on the estimated slope and its standard error suggests that we can with high confidence reject the null hypothesis that there is no general tendency for the mean atrophy scores to differ in a linear fashion across age groups (P < .0005). Based on the 95% CI, we can with confidence state that the average ratio between geometric mean atrophy for two groups differing in age by 1 year is between a 1.46% and a 2.42% higher geometric mean per year difference in age. That is, the observed estimate of a 1.92% higher geometric mean per year difference in age is not atypical of what we might expect to see if the true geometric mean ratio were between 1.0146 and 1.0242.

c. Using the regression model estimated in your analysis, estimate the geometric mean atrophy score for 70 year olds and for 90 year olds. Do you believe that such estimates would accurately reflect the true geometric mean atrophy score in the populations of all 70 or 90 year olds? Why or why not?

Answer: From the regression model, we estimate the geometric mean atrophy scores for 70 year olds as exp(2.079 + 70 * .0192) = 30.7, and we estimate the geometric mean atrophy scores for 90 year olds patients as exp(2.079 + 90 * .0192) = 45.2. From a lowess curve of the plot of log atrophy scores versus age, I do not see marked departures from linearity, thus I would tend to believe that the above model produces reasonable approximations.

4. Perform an analysis to determine whether there is a trend across age groups in the tendency for cerebral atrophy scores to be greater than 30. (Note: The Stata command "logit" provides output for both the slope and the intercept on the log odds ratio scale. The Stat command "logistic" provides output only for the slope on the odds ratio scale. Usually the "logistic" output will suffice, but for parts a and c you will probably want to use the "logit" command.)

```
. g atrgt30 = atrophy
. recode atrgt30 0/30=0 30/max=1
. logit atrgt30 age, robust
Logistic regression                                     Number of obs   =        735
                                                        Wald chi2(1)    =      34.96
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -460.38933                       Pseudo R2       =     0.0395
------------------------------------------------------------------------------
             |               Robust
     atrgt30 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0952247   .0161063     5.91   0.000      .063657    .1267924
       _cons |  -6.473034   1.189368    -5.44   0.000    -8.804153   -4.141915
------------------------------------------------------------------------------

. logistic atrgt30 age, robust
Logistic regression                                     Number of obs   =        735
                                                        Wald chi2(1)    =      34.96
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -460.38933                       Pseudo R2       =     0.0395
------------------------------------------------------------------------------
             |               Robust
     atrgt30 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.099906   .0177154     5.91   0.000     1.065727    1.135181
------------------------------------------------------------------------------
```

a. Provide an interpretation of the intercept from the regression model, along with statistical inference pertaining to the hypothesis that the intercept of the regression model might be 0. What is the scientific relevance of the answer to this question?

Answer: From a logistic regression model, we estimate an intercept of -6.473, from which we estimate that the odds of newborns (age 0) having an atrophy score greater than 30 is 0.00154 (by exponeniating the intercept), which corresponds to a probability of 0.00154 (probability = odds / (1 + odds)). Testing that the intercept is 0 is equivalent to testing that the odds of high atrophy (greater than 30) in newborns would be 1. The Z test based on the estimated intercept and its standard error suggests that we can with high confidence reject the null hypothesis of a 50% probability (odds of 1.0) that a newborn would have an atrophy score greater than 30. However, given that we had no patients under 65 in our study, we should be extremely reluctant to use this data to try to estimate what the distribution of atrophy scores would be in newborns. So the intercept is not of scientific interest for its own sake.

b. Provide an interpretation of the slope from the regression model, along with statistical inference pertaining to the hypothesis that the slope might be 0. What is the scientific relevance of the answer to this question?

Answer: From the logistic regression model, we estimate a slope of 0.0952 from which we estimate that the odds ratio comparing the distribution of atrophy scores greater than 30 between two groups of older adults that differ in age by 1 year is 1.0999 (found by exponentiation). Thus, when comparing groups that differ in age

by 1 year, we estimate that the odds of high atrophy scores is 9.99% higher in the older group. The Z test based on the estimated slope and its standard error suggests that we can with high confidence reject the null hypothesis that there is no general tendency for the odds of high atrophy scores to differ across age groups (P < .0005). Based on the 95% CI, we can with confidence state that over the ages sampled in this data, the odds ratio comparing two groups differing in age by one year is between 1.066 and 1.135.

c. Using the regression model estimated in your analysis, estimate the proportion of 70 year olds and 90 year olds who would have a cerebral atrophy score greater than 30. Do you believe that such estimates would accurately reflect the true proportions in the populations of all 70 or 90 year olds? Why or why not?

Answer: From the regression model, we estimate the odds of high atrophy scores for 70 year olds as exp(-6.473 + 70 * 0.0952) = 1.210, from which we estimate the probability (= odds / (1+odds)) of high atrophy scores as 0.548. For 90 year olds the odds of high atrophy is 8.125, corresponding to a probability of 0.89. Assessing the linearity of the log odds of high atrophy as a function of age is difficult. One approach would be to calculate the log (p / (1-p)) for the last column in Table 1. When I did that, I found that the trend seemed linear over the first few categories, with some hint of leveling off in the highest categories (but the sample size was quite small). We could also try fitting a logistic regression model with age and, say, age squared to see if there was a statistically significant departure from linearity that could be modeled by a quadratic. When I did this, I found that the age squared term's coefficient was not statistically significant. This, of course, only means that I could not prove the trend in log odds was nonlinear. I do note that the estimates I obtained from the logistic regression model were not that different from the descriptive statistics shown in Table 1 for those strata.

Questions 5 - 6 investigate whether the distribution of survival time differs by atrophy level in the population from which the MRI measurements were sampled.

5. Provide suitable descriptive statistics related to the distribution of survival times by atrophy level.

Answer: Figure 2 displays survival curve estimates within strata defined by atrophy divided into categories of width 15. Table 2 presents estimates of survival at 1, 2, 5, and 10 years within those same strata. Immediately apparent is the decreased probability of survival in groups with higher bilirubin levels. It should be noted that successive bilirubin strata represent approximate doubling of the baseline bilirubin levels.

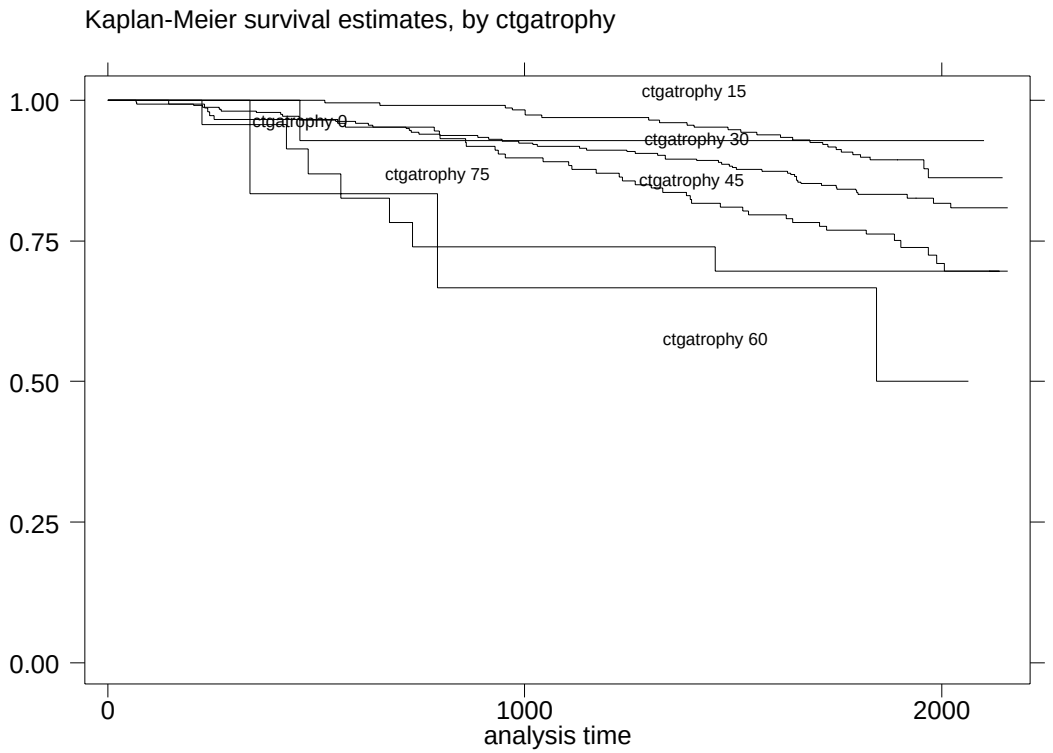

Figure 2: Kaplan-Meier estimates of survival curves within strata defined by atrophy scores.

Table 2: Survival probabilities within atrophy strata estimated by Kaplan-Meier curves.

	Survival Probabilities		
	1 Year	2 Year	5 Year
Atrophy 0 - 14	1.000	0.929	0.929
Atrophy 15 – 29	1.000	0.992	0.899
Atrophy 30 – 44	0.978	0.943	0.833
Atrophy 45 – 59	0.966	0.952	0.762
Atrophy 60 – 74	0.957	0.783	0.696
Atrophy 75 - 89	0.833	0.833	0.667

6. Perform an analysis to determine whether there is a trend in survival time distribution across groups defined by atrophy level. (Note: The Stata command “stcox” will perform proportional hazards regression for the survival variables declared in a “stset” command. The output will be on the hazard ratio scale.)

```
. stset obstime death
. stcox atrophy, robust

        failure _d:  death
  analysis time _t:  obstime

Cox regression -- Breslow method for ties

No. of subjects      =          735        Number of obs    =       735
No. of failures      =          133
Time at risk         =      1325995
                                           Wald chi2(1)     =     23.49
```

```
Log pseudolikelihood =   -844.04581          Prob > chi2     =    0.0000
------------------------------------------------------------------------------
             |               Robust
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     atrophy |   1.030879   .0064684     4.85   0.000     1.018278    1.043635
------------------------------------------------------------------------------
```

a. Provide an interpretation of the slope from the regression model, along with statistical inference pertaining to the hypothesis that the slope might be 0. What is the scientific relevance of the answer to this question?

Answer: From the proportional hazards regression model, we estimate a hazard ratio comparing the survival distribution between two groups of older adults that differ in atrophy scores by 1 unit is 1.031. Thus, when comparing groups that differ in atrophy by 1, we estimate that the instantaneous risk of death is 3.1% higher in the group with the higher atrophy score. The Z test based on the estimated slope and its standard error suggests that we can with high confidence reject the null hypothesis that there is no general tendency for the survival distribution to differ across groups defined by atrophy score (P < .0005). Based on the 95% CI, we can with confidence state that over the atrophy levels sampled in this data, the hazard ratio comparing two groups differing in atrophy by 1 is between 1.018 and 1.044. Note that for every 10 unit difference in atrophy scores, we estimate a hazard ratio of $1.031^{10} = 1.355$, or a 35.5% higher risk of death in the subjects with higher atrophy scores.