**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2014

**Homework #5**

February 3, 2014

**Written problems:** To be submitted as a MS-Word compatible file to the class Catalyst dropbox by 9:30 am on Monday, February 10, 2014. See the instructions for peer grading of the homework that are posted on the web pages.

*On this (as all homeworks) Stata / R code and unedited Stata / R output is* ***TOTALLY*** *unacceptable. Instead, prepare a table of statistics gleaned from the Stata output. The table should be appropriate for inclusion in a scientific report, with all statistics rounded to a reasonable number of significant digits. (I am interested in how statistics are used to answer the scientific question.)*

***Unless explicitly told otherwise in the statement of the problem, in all problems requesting “statistical analyses” (either descriptive or inferential), you should present both***

* ***Methods: A brief sentence or paragraph describing the statistical methods you used. This should be using wording suitable for a scientific journal, though it might be a little more detailed. A reader should be able to reproduce your analysis. DO NOT PROVIDE Stata OR R CODE.***
* ***Inference: A paragraph providing full statistical inference in answer to the question. Please see the supplementary document relating to “Reporting Associations” for details.***

Problems 2 and 3 of the homework build on the analyses performed in homeworks #1 through #4. As such, all questions relate to associations among death from any cause, serum low density lipoprotein (LDL) levels, age, and sex in a population of generally healthy elderly subjects in four U.S. communities. This homework uses the subset of information that was collected to examine MRI changes in the brain. The data can be found on the class web page (follow the link to Datasets) in the file labeled mri.txt. Documentation is in the file mri.pdf. See homework #1 for additional information. Problem 1 of this homework uses the same dataset to explore associations between prevalence of diabetes and race in the population from which that sample was drawn.

1. Perform a statistical regression analysis evaluating an association between prevalence of diabetes and race by comparing the odds of a diabetes diagnosis across.
	1. Fit a logistic regression model that uses whites as a reference group. Is this a saturated model? Provide a formal report (methods and inference) about the scientific question regarding an association between diabetes and race.

**Answer:** *Since we are fitting a model that has 2 distinct outcome groups (presence/absence of diabetes) and 4 parameters (based on 4 categories of race) we do not have a saturated model (number of outcome groups ≠ number of parameters).*

**Methods: The presence of an association between race and prevalence of diabetes was tested using model comparison via the Wald test based on the F-statistics, where the an unadjusted model of diabetes prevalence was compared to a logistic regression model which included the estimated odds of subjects having diabetes, based on the predictor race.**

**Among all 735 subjects, 89.3% (656) were not observed to have diabetes and 10.7% (79) were observed to have the disease; and 77.8% (572) were white, 14.2% (104) were black, 6.4% (47) were Asian, and 1.6% were of other races. These results are summarized in Table 1.**

Table 1

|  |
| --- |
| **Frequency table of Diabetes based on Race** |
| **Race** | **N** | **# Diabetic** | **# Non-diabetic** | **Percent of total** |
| **White** | 572 | 56 | 516 | 77.8 |
| **Black** | 104 | 18 | 86 | 14.2 |
| **Asian** | 47 | 3 | 44 | 6.4 |
| **Other** | 12 | 2 | 10 | 1.6 |

**Results: Based on the F statistic from a Wald test comparing the fit of the models, we conclude that there is not a significant association between diabetes and race, based on a 2 sided p value 0.09626.**

* 1. Using the regression model fit in part (a), provide an interpretation for each of the regression parameters (including the intercept).

**The intercept parameter represents the odds of having diabetes for white subjects, compared to the rest of the study population. The interpretation of the first, second, and third slope parameters is the odds ratio of having diabetes for black, Asian, and other-raced subjects, respectively, relative to white subjects. Since there were no other predictors besides race in this model, these parameter estimates correspond exactly with the odds ratios observable within the sample.**

* 1. If we were to ignore issue related to multiple comparisons, what conclusions would you reach based on the p values reported in the regression output from part (a) using a 0.05 level of significance.

**Contrary to the results in part (a), we conclude that there is a significant association between diabetes and race. Based on the logistic regression model using white subjects as a reference group, the odds of having diabetes was 92.8% higher for black subjects, when compared to white subjects (two-sided p value 0.0259) (diabetes odds ratio = 1.93). Based on a 95% confidence interval, this observed odds ratio suggesting higher diabetes probability for black subjects would not be judged unusual if the true odds ratio were anywhere from 8.19% to 245% higher odds of diabetes compared to white subjects (95% CI for diabetes odds ratio 1.082 to 3.438). Asian and other races were not observed to have significantly different odds of diabetes, compared to white subjects (95% CI for diabetes odds ratio comparing Asian subjects to white subjects: 0.189 to 2.089, two-sided p value: 0.448. 95% CI for diabetes odds ratio comparing other-race subjects to white subjects: 0.3934 to 8.622, two sided p value 0.438).**

* 1. Now fit a logistic regression model that uses blacks as a reference group. How would your report of formal inference differ from that that you provided in part (a)? How does this regression model relate to that in part (a)?

**The inference is exactly identical to part (a), since both analyses are comparing the null model to a model fitted with race. The only difference is that the new logistic regression model being compared to the unadjusted null model is a re-parameterization of that of part (a). The regression model yields estimates that are consistent with part (a) (that is we can compute the sample odds of diabetes for any given race based on these estimates), but using black subjects as a reference group, and comparing the odds of diabetes among all other races (white, Asian and other) to that of black subjects.**

* 1. Using the regression model fit in part (d), provide an interpretation for each of the regression parameters (including the intercept.)

**The intercept parameter represents the odds of having diabetes for black subjects, compared to the rest of the study population. The interpretation of the first, second, and third slope parameters is the odds ratio of having diabetes for Asian, other-raced, and white subjects, respectively, relative to black subjects. Since there were no other predictors besides race in this model, these parameter estimates correspond exactly with the odds ratios observable within the sample.**

* 1. If we were to ignore issue related to multiple comparisons, what conclusions would you reach based on the p values reported in the regression output from part (d) using a 0.05 level of significance.

**As in the regression model in part (a), here we do observe a significantly different odds of diabetes when we compare white subjects to our reference group (black subjects), and obtain the same p value since the comparison is again between black and white subjects. Based on the logistic regression model using black subjects as a reference group, the odds of having diabetes was 42.1% lower for white subjects, when compared to black subjects (two-sided p value 0.0259) (diabetes odds ratio = 0.518). Based on a 95% confidence interval, this observed odds ratio suggesting lower diabetes probability for white subjects would not be judged unusual if the true odds ratio were anywhere from 7.57% to 71.9% lower odds of diabetes compared to black subjects (95% odds ratio CI 0.291 to 0.924). The estimated odds ratio and 95% confidence interval for white subjects compared to black subjects is just the inverse of that obtained from part (c). Because we are now using black subjects as a reference group, the other odds ratios and p values (i.e. comparing odds of diabetes among Asians and other races compared to black subjects) are different from that of part (a). However, we still do not observe a statistically significant difference in odds of diabetes (two-sided p values 0.0847 and 0.9556 for odds ratios comparing Asian and other races to that black subjects, respectively).**

* 1. What do your results from parts (c) and (f) say about the dangers of using the p values for individual regression parameters from a dummy variable regression to decide whether to include or exclude those variables in a regression model (i.e., in a “stepwise model building” procedure)?

**We observe that although there were significant p values for individual regression parameters, the Wald test comparing the null and adjusted models did not provide evidence of an overall significant association between race and diabetes. The Wald test of association appropriately provides a more conservative test than that of each regression parameter in the logistic regression model.**

**Due to the fact that logistic regression involving more than 2 parameters inflates our probability of falsely rejecting the null hypothesis the Wald test is a better analysis of association in this case.**

1. Perform a statistical regression analysis evaluating an association between all-cause mortality and serum by comparing the instantaneous risk (hazard) of death over the entire period of observation across groups defined by serum LDL when fit as dummy variables using the categories suggested by the Mayo Clinic as reported on Homework #1. The Stata egen command can be used to categorize the LDL levels

egen ldlCTG = cut(ldl), at(0 70 100 130 160 190 250)

* 1. Include full description of your methods, appropriate descriptive statistics, and full report of your inferential statistics.

**Methods for Descriptive statistics: We created a graph and a table depicting the Kaplan-Meier estimates of survival probability for subjects defined by blood serum LDL levels, categorized according to the Mayo Clinic guidelines: less than 70 mg/dL, 70-99 mg/dL, 100-129 mg/dL, 130-159 mg/dL, 160-189 mg/dL, and greater than or equal to 190 mg/dL. Within these categories, Kaplan-Meier estimates of survival were calculated and graphed, and estimates of the 2 and 5 year survival probabilities, as well as restricted mean survival during a period of observation that all LDL strata still had some subjects at risk (5.75 years).**

**Descriptive statistics: The study consisted of 735 subjects who were followed for death from any cause, 10 of whom were omitted from the analysis due to missing information on baseline serum LDL levels, for a total of 725. Among the 725 subjects with available serum LDL measurements at enrollment, the mean serum LDL was 126 mg/dL (SD 33.6 mg/dL) with values ranging from 11 to 247 mg/dL.**

**Table 1 presents survival data on groups defined by serum LDL levels, as well as the entire group of study participants. Evident from Table is that 2 year survival is highest (100% survival rate) among those subjects with the lowest serum LDL (0-69 mg/dL) and lowest (93.9% survival rate) among subjects with an intermediate serum LDL level of 100-129 mg/dL. Contrary to this trend, we find 5 year survival is markedly lower in the 0-69mg/dL serum LDL group, while all other groups range between 81%-88% 5 year survival probability, with the higher LDL group of 159-190 mg/dL having the greatest survival probability. This 5 year survival trend is also evident from the 5.75 year restricted mean survival, which measures the area under the Kaplan-Meier curve (Figure 1), across the range of time for which we have survivors in every level of serum LDL.**

**Figure 1 shows us that among all 6 levels of serum LDL, we observe similar survival curves for all levels besides 0-69mg/dL, whose survival curve deviates from the pattern of the other curves around 3.5 years, in the direction of lower survival.**

|  |
| --- |
| **Survival statistics based on Serum LDL at Study Enrollment** |
|  | **0-69 mg/dL** | **70-99 mg/dL** | **100-129 mg/dL** | **130-159 mg/dL** | **159-190 mg/dL** | **>190 mg/dL** | **All levels** |
| **Number** | 22 | 143 | 228 | 225 | 83 | 24 | 725 |
| **2 yr survival** | 100% | 95.8% | 93.9% | 95.6% | 98.8% | 95.8% | 95.6% |
| **5 yr survival** | 59.1% | 83.2% | 81.1% | 87.1% | 88.0% | 83.3% | 83.6% |
| **5.75 year Restricted mean survival** | 4.91 years | 5.24 years | 5.23 years | 5.35 years | 5.45 years | 5.32 years | 5.29 years |



**Methods for inferential Statistics: We estimated presence of an association between survival and blood serum LDL using the Wald test and chi-square statistic, comparing an unadjusted model to the cox proportional hazard regression model based on the ordinal categorical predictor serum LDL (low density lipoprotein, measured in mg/dL). LDL was split into the following categories, defined by the Mayo Clinic guidelines: less than 70 mg/dL, 70-99 mg/dL, 100-129 mg/dL, 130-159 mg/dL, 160-189 mg/dL, and greater than or equal to 190 mg/dL.**

**Results: Based on this analysis, we conclude that there is a significant association between serum LDL and survival probability (two-sided p value 0.0087)**

* 1. Provide an interpretation for each parameter in your regression model, including the intercept.

**The intercept of this model is the estimated instantaneous risk of death for the group of subjects whose blood serum LDL is lowest (between 0-69 mg/dL, this group is used as the reference for the other 5 LDL groups). Each of the other parameters represent the ratio of the estimated instantaneous risk of death for each category of serum LDL (other than the reference group) compared to the reference group of 0-69 mg/dL.**

* 1. What analysis would you perform to assess whether the regression model used in this problem provides a “better fit” than does a model that uses only a continuous linear term for LDL? What is the result of such an analysis?

**To test the fit efficacy of the model using 6 categories of LDL, compared to the model using a continuous linear predictor of LDL we use the Wald test based on the chi square-statistic, where two proportional hazards regression models are compared: the one which included the continuous linear predictor LDL and the other having this predictor in addition to the 6 categories of LDL. Based on this test, we fail to reject the hypothesis that adding a categorical term provides a better fit than the linear model, based on a two-sided p value of 0.3988. Conducting this analysis specifically tells us that there is not sufficient evidence for non-linearity in survival probability, as it relates to serum LDL, based on the stratification that we employ.**

* 1. For each population defined by serum LDL value, compute the hazard ratio relative to a group having serum LDL of 160 mg/dL. (This will be used in problem 4). This can be effected by generating fitted hazard ratio estimates for each individual in the sample, and then dividing that fitted value by the fitted value for a subject having a LDL of 160 mg/dL.
1. Perform a statistical regression analysis evaluating an association between all-cause mortality and serum by comparing the instantaneous risk (hazard) of death over the entire period of observation across groups defined by serum LDL when fit as linear splines using the categories suggested by the Mayo Clinic as reported on Homework #1. The Stata mkspline command can be used to create the predictors that can be used in a regression

mkspline ldl0 70 ldl70 100 ldl100 130 ldl130 160 ldl160 190 ldl190 = ldl

* 1. Include full description of your methods, appropriate descriptive statistics, and full report of your inferential statistics.

**Methods and inference for descriptive Statistics: (See 2a)**

**Methods for inferential Statistics: We estimated presence of an association between survival and blood serum LDL using the Wald test comparing an unadjusted model to the cox proportional hazard regression model based on the predictor serum LDL (low density lipoprotein, measured in mg/dL), fit by linear splines using the following intervals defined by the Mayo Clinic guidelines: less than 70 mg/dL, 70-99 mg/dL, 100-129 mg/dL, 130-159 mg/dL, 160-189 mg/dL, and greater than or equal to 190 mg/dL.**

**Results: Based on this analysis, we conclude that there is a significant association between serum LDL and survival probability (two-sided p value <** **0.0001)**

* 1. Provide an interpretation for each parameter in your regression model, including the intercept.

**The intercept is the instantaneous risk of death for someone having a blood serum LDL level of 0, and so is not scientifically relevant, since is not in the range of our data. Each parameter represents the average change in instantaneous risk of death when serum LDL is increased by 1.0 mg/dL, within the interval defined by that parameter (e.g. the first parameter corresponds to the average change in instantaneous risk of death when serum LDL is increased by 1 unit among subjects having serum LDL in the range of 0-69 mg/dL).**

* 1. What analysis would you perform to assess whether the regression model used in this problem provides a “better fit” than does a model that uses only a continuous linear term for LDL? What is the result of such an analysis?

**To test the fit efficacy of the model using linear splines of LDL (based on the predefined intervals) compared to the model using a continuous linear predictor of LDL we use the Wald test based on the chi squared-statistic, where two proportional hazards regression models are compared: the one which includes a continuous linear predictor LDL and the other having this predictor in addition to the spline estimates for the 6 categories of LDL. Based on this test, we fail to reject the hypothesis that adding a categorical term provides a better fit than the linear model, based on a two-sided p value of 0.0788. Conducting this analysis specifically tells us that there is not sufficient evidence for non-linearity in survival probability, as it relates to serum LDL, based on using linear splines for each subcategory.**

* 1. For each population defined by serum LDL value, compute the hazard ratio relative to a group having serum LDL of 160 mg/dL. (This will be used in problem 4). This can be effected by generating fitted hazard ratio estimates for each individual in the sample, and then dividing that fitted value by the fitted value for a subject having a LDL of 160 mg/dL.
1. By answering the following questions, compare the relative advantages and disadvantages of the various statistical analysis strategies we have considered in Homeworks 1-4 and problems 2 and 3 in this homework.
	1. What advantages do the regression strategies used in Homeworks 4 and 5 provide over the approaches used in Homeworks 1-3?

**The strategies in homeworks 4 and 5 allow for analysis when trends in the data are non-linear (higher order, log-linear) and non-parametric (piece-wise linear). This gives the potential to test for these abnormalities, and also may provide a greater precision due to better fit (this also applied to measuring effects of confounders).**

* 1. Comment on any similarities or differences of the fitted values from the three models fit in Homework 4 and the two models fit in problems 2 and 3 of this homework.

**The graph below shows the estimated relative hazard ratio for each value of serum LDL compared to someone having serum LDL of 160 mg/dL. Compared to the fits in homework 4, the linear spline model gives a slightly lower estimation of relative HR compared to the 160 mg/dL referent, though it agrees with these models a bit more in the tails of the data. The categorical model shows more or less the average value of the models in homework 4, within each stratum created. This model agrees with the overall trend observed in the 3 models from homework 4; however it provides lower accuracy since it takes the average HR across intervals of serum LDL.**



* 1. *A priori*, of all the analyses we have considered for exploring an (unadjusted) association between all cause mortality and serum LDL in an elderly population, which one would you prefer and why?

***A priori*, I would hypothesize an association (if it exists) between blood serum LDL and mortality to have a linear or log-linear (multiplicative) effect on mortality, since previous literature suggests that increased serum LDL tends to increase the risk of fatal illness and disease, and thereby increase the risk of death. Since we do not know *a priori* whether or not LDL might have a multiplicative effect on survival, I would prefer to conduct a cox proportional hazards regression analysis of survival on serum LDL modeled as a continuous predictor.**

**Discussion Sections: February 3 - 7, 2014**

We continue to discuss the dataset regarding FEV and smoking in children. Come do discussion section prepared to describe descriptive statistics, especially as they relate to confounding, precision, effect modification, and the impact of heteroscedasticity.