**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.
		1. **Since the model observes 4 sub-groups while containing only 4 parameters, the model is saturated.**
		2. **Method: The odds of diagnosing subjects with diabetes were compared across race membership using a logistic regression model. Memberships considered in this analysis include “White,” “Black,” “Asian,” and “Other.” Important to note that “White” will be used as the reference group. Statistical inference was based on the Wald statistic computed from the regression coefficients per each membership and their standard errors, with two-sided p value and 95% confidence interval computed using the approximate normal distribution for logistic regression parameter estimates. Furthermore, association between diabetes and race will be evaluated using the overall F-test.**
		3. **Results: Inference regarding the association between diabetes and specific race is not of interest in this question. Instead, we are interested in the general association between diabetes and race without focus. At the 5% significance level, we fail to reject the null hypothesis for no association between diabetes and race (P = 0.0963).**
	2. Using the regression model fit in part (a), provide an interpretation for each of the regression parameters (including the intercept).
		1. **Methods: Please see part (a) for the methods used in this problem.**
		2. **Results: The logistic regression model considered in part (a) yields an Intercept of -2.22, which after exponentiation translates to an odds ratio of 0.109. Hence, since we have a saturated model, 0.109 corresponds to the sample odds of a “White” person being diagnosed with diabetes. Now for each subsequent parameter, we now interpret the coefficients as odds ratios. Persons with “Black” race membership, with estimated coefficient 0.657, have 92.9% greater odds of being diagnosed with diabetes than those under “White” race membership. Persons with “Asian” race membership, with estimated coefficient -0.4648, have 37.2% lesser odds of being diagnosed with diabetes in comparison to the “White” odds of diabetes diagnosis. Finally, persons with “Other” membership, with estimated coefficient 0.6113, have 84.3% greater odds of being diagnosed with diabetes than those with “White” race membership.**
	3. 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.
		1. **Methods: Please see part (a) for the methods used in this problem.**
		2. **Results: Based only on the p-values reported in the regression output for each parameter, the conclusion would vary from each race membership, although each would be compared against “White.” At the 5% significance level, we would reject the null hypothesis of no association between diabetes and race when comparing “Black” against “White” (P=0.026). However, at the 5% significance level, we would fail to reject the null hypothesis of no association between diabetes and race when comparing “Asian” against “White” (P=0.448). Similarly, we would fail to reject the null hypothesis of no association between diabetes and race when comparing “Other” against “White” (P=0.438).**
	4. 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)?
		1. **Method: The odds of diagnosing subjects with diabetes were compared across race membership using a logistic regression model. Memberships considered in this analysis include “White,” “Black,” “Asian,” and “Other.” Important to note that “Black” will be used as the reference group. Statistical inference was based on the Wald statistic computed from the regression coefficients per each membership and their standard errors, with two-sided p value and 95% confidence interval computed using the approximate normal distribution for logistic regression parameter estimates. Furthermore, association between diabetes and race will be evaluated using the overall F-test.**
		2. **Results: Concerning the association between diabetes and race in “general,” the report of formal of formal inference would be the same as in part (a): At the 5% significance level, we fail to reject the null hypothesis for no association between diabetes and race (P = 0.0963). The general association is being tested with the Overall F-test, which will be the same test in both models considered. The main difference between these models is the reference group. Before we contrasted race membership against “White,” and now we contrast with “Black.” Hence, our across group inference remains the same, while our within group inference changes.**
	5. Using the regression model fit in part (d), provide an interpretation for each of the regression parameters (including the intercept.)
		1. **Methods: Please see part (c) for the methods used in this problem.**
		2. **Results: The logistic regression model considered in part (a) yields an Intercept of -1.56, which after exponentiation translates to an odds of 0.209. Hence, since we have a saturated model, 0.209 corresponds to the sample odds of a “Black” person being diagnosed with diabetes. For each subsequent parameter, we now interpret the coefficients as odds ratios. Persons with “White” race membership, with estimated coefficient -0.657, have 48.1% lower odds of being diagnosed with diabetes than those under “Black” race membership. Persons with “Asian” race membership, with estimated coefficient -1.122, have 67.4% lower odds of being diagnosed with diabetes in comparison to the “Black” odds of diabetes diagnosis. Finally, persons with “Other” membership, with estimated coefficient -0.046, have 4.5% lesser odds of being diagnosed with diabetes than those with “Black” race membership.**
	6. 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.
		1. **Methods: Please see part (c) for the methods used in this problem.**
		2. **Results: Based only on the p-values reported in the regression output for each parameter, the conclusion would vary from each race membership, although each would be compared against “Black.” At the 5% significance level, we would reject the null hypothesis of no association between diabetes and race when comparing “White” against “Black” (P=0.026). However, at the 5% significance level, we would fail to reject the null hypothesis of no association between diabetes and race when comparing “Asian” against “Black” (P=0.085). Similarly, we would fail to reject the null hypothesis of no association between diabetes and race when comparing “Other” against “Black” (P=0.956).**
	7. 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)?
		1. **Although we may observe a significant p-value amongst the individual parameters and that will be in general rather consistent with the overall association between the response (diabetes) and the predictor (race) being tested, caution must be drawn since that association may be specific to the reference group, but cannot be generalizable outside of that comparison. The individual p-values can change as we change the reference, so it becomes important to consider a stepwise model building approach where we cycle through each category.**
2. 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.
		1. **Method: Descriptive statistics for serum LDL levels included the number of cases with missing data, as well as the minimum, maximum, mean, and standard deviation for the cases with available data. For the purposes of descriptive statistics of the survival probabilities by serum LDL level, serum LDL was 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 the 10th and 20th percentiles of the survival distribution and the restricted mean survival during a period of observation that all LDL strata still had some subjects at risk (5.75 years).**
		2. **Distributions of time to death from any cause was compared across groups defined by serum LDL using proportional hazards regression modeling serum LDL as a categorized (by Mayo clinic guidelines) random variable. Note that the model will use those with LDL below 70 as reference group. Quantification of associations between mortality and category of LDL was summarized by the hazards ratios computed from the regression model, with confidence intervals and two-sided p values computed using Wald statistics based on the Huber-White sandwich estimator. Subjects missing data for serum LDL at the time of study accrual were omitted from the analysis. Using an overall F test at the 5% significance level, we evaluate the general association between LDL and mortality.**
		3. **Results:** **The study contains 735 patients followed for any cause death.  Ten patients were missing LDL data and were therefore excluded from this analysis.  The table below shows estimates for survival for strata defined by LDL levels for the 725 patients with recorded LDL measurements.  In the 725 subjects with available serum LDL measurements at enrollment, the mean LDL was 126 mg/dL (SD 33.6 mg/dL, range 11 to 247 mg/dL). On average, the subjects in the lowest LDL stratum were estimated to average 4.91 years of life during the first 5.75 years following study enrollment, while the other strata averaged from 5.23 to 5.45 years.  The lowest strata had much lower 5 year survival (59.1%) than the other five strata (survival ranging from 81.1% to 88.0%), despite having the highest 2 year survival (100%).  The Kaplan-Meier plot makes the low survival rate for the lowest LDL group even more apparent.**
		4. **Using a Cox proportional hazards regression model, we evaluate the associations between categorized LDL and mortality. At the 5% significance level, we reject the null hypothesis of no association between LDL and mortality (P=0.0086).**





* 1. Provide an interpretation for each parameter in your regression model, including the intercept.
		1. **Methods: Please refer to the methods in part (a) (problem 2, naturally).**
		2. **Results: Since we have a factorized LDL variable in the Cox proportional hazards regression model, a factor is being used as reference for the hazard ratios. In this model, those with serum LDL at or below 70 mg/dL are the reference. Hence, the intercept would be the ratio of the hazard for subjects with serum LDL at or below 70 mg/dL to themselves, which is 1 by default. We gain no information from this kind of baseline hazard. For those with LDL between 70 and 100 mg/dL, we observe a regression parameter of (once exponentiated) 0.3981. Thus, from the proportional hazards model, those subjects with LDL between 70 and 99 mg/dL have 60.19% lower instantaneous risk of death than those subjects with LDL below 70 mg/dL. Similarly, interpretations can be given for the other parameters after exponentiating. From the model, those with LDL between 100 and 129 mg/dL have 60.74% lower instantaneous risk of death in comparison to subjects with LDL below 70 mg/dL. Subjects with LDL between 130 and 159 mg/dL have an instantaneous risk of death that is 70.61% smaller than that of subjects with LDL below 70 mg/dL. Subjects with LDL between 160 and 189 mg/dL have 74.34% lower instantaneous risk of death than subjects with LDL below 70 mg/dL. Finally, from the model, subjects with LDL between 190 and 247 mg/dL have 68.33% lower instantaneous risk of death than subjects with LDL below 70 mg/dL.**
	2. 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?
		1. **Method: To assess whether the regression model used in this problem provides better fit than does a model that uses only a continuous linear term for LDL, we first create a Cox proportional hazards regression model based only on the continuous LDL predictor. Using a likelihood ratio test, the models can be simultaneously tested against each other with respect to their likelihoods, which serve as a measure for their fit in this problem.**
		2. **Results: At the 5% significance level, we fail to reject the null hypothesis that using a categorical LDL term will result in a “better fitting” model (P=0.478). We are unable to conclude that the model with continuous LDL will have a better fit than the model with a continuous LDL.**
	3. 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. **Problem 4 will have the values.**
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.
		1. **Method: Descriptive statistics for serum LDL levels included the number of cases with missing data, as well as the minimum, maximum, mean, and standard deviation for the cases with available data. For the purposes of descriptive statistics of the survival probabilities by serum LDL level, serum LDL was 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 the 10th and 20th percentiles of the survival distribution and the restricted mean survival during a period of observation that all LDL strata still had some subjects at risk (5.75 years).**
		2. **Distributions of time to death from any cause was compared across groups defined by serum LDL using proportional hazards regression modeling serum LDL with linear splines (knots defined by Mayo clinic guidelines). Note that the model will use those with LDL below 70 as reference group. Quantification of associations between mortality and category of LDL was summarized by the hazards ratios computed from the regression model, with confidence intervals and two-sided p values computed using Wald statistics based on the Huber-White sandwich estimator. Subjects missing data for serum LDL at the time of study accrual were omitted from the analysis. Using an overall F test at the 5% significance level, we evaluate the general association between LDL and mortality.**
		3. **Results:** **The study contains 735 patients followed for any cause death.  Ten patients were missing LDL data and were therefore excluded from this analysis.  The table below shows estimates for survival for strata defined by LDL levels for the 725 patients with recorded LDL measurements.  In the 725 subjects with available serum LDL measurements at enrollment, the mean LDL was 126 mg/dL (SD 33.6 mg/dL, range 11 to 247 mg/dL). On average, the subjects in the lowest LDL stratum were estimated to average 4.91 years of life during the first 5.75 years following study enrollment, while the other strata averaged from 5.23 to 5.45 years.  The lowest strata had much lower 5 year survival (59.1%) than the other five strata (survival ranging from 81.1% to 88.0%), despite having the highest 2 year survival (100%).  The Kaplan-Meier plot makes the low survival rate for the lowest LDL group even more apparent.**
		4. **Using a Cox proportional hazards regression model with LDL linear splines, we evaluate the associations between LDL and mortality. At the 5% significance level, we reject the null hypothesis of no association between LDL and mortality (P<0.0001).**





* 1. Provide an interpretation for each parameter in your regression model, including the intercept.
		1. **The parameters for these splines are interpreted within the group they represent. Within subjects with LDL less than 70 mg/dL, per 1 mg/dL increase, the subject with higher LDL has a risk of instantaneous death that is 2.2% lower than the subject with less LDL. Within subjects with LDL at or higher than 70 but below 100 mg/dL, per 1 mg/dL increase, the subject with higher LDL has a risk of instantaneous death that is 2.0% lower than the subject with less LDL. Within subjects with LDL at or higher than 100 but below 130 mg/dL, per 1 mg/dL increase, , the subject with higher LDL has a risk of instantaneous death that is 0.20% lower than the subject with less LDL. Within subjects with LDL at or higher than 130 but below 160 mg/dL, per 1 mg/dL increase, the subject with higher LDL has a risk of instantaneous death that is equal to the subject’s risk with less LDL. Within subjects with LDL at or higher than 160 mg/dL, the subject with higher LDL has a risk of instantaneous death that is 2.9% lower than the subject with less LDL. Finally, within subjects presenting with LDL at or greater than 190 but less than 248 mg/dL, for each 1 mg/dL increase in LDL, the subjects with greater LDL has 3% greater risk of instantaneous death than those with lower LDL.**
	2. 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?
		1. **Method: Using a likelihood ratio test, the models can be simultaneously tested against each other with respect to their likelihoods, which serve as a measure for their fit in this problem.**
		2. **Results: At the 5% significance level, we fail to reject the null hypothesis that using a continuous LDL is as fitting as a linear spline LDL term model (P=0.373). We do not have sufficient evidence to support the claim that the model with linear splines of LDL has a better fit than the model with continuous LDL.**
	3. 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. Please see 4.
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?
		1. **The greatest advantage gained in Homework 4 and 5 is the ability to regress over time. In previous homework’s, our response of mortality was dichotomized in such a way that we were forced to consider mortality at 5 years. Using Cox proportional hazards regression, we have been able to consider the relationship between LDL and death over all time, not just at a prescribed cut-off like 5 years. Furthermore, our regression strategies have evolved so that we no longer have to consider dichotomous LDL: we can choose it to be categorical over 5 or 6 levels and we can even improve the fit of the model by creating linear splines.**
	2. 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.
		1. **All of the fitted curves seem to follow the same trend in direction. The spline model is working harder to estimate the outlier on the far top left of the graph, which may lead to bloated estimates for LDL values less than 100 mg/dL in that model. However, the spline estimates would serve estimating purposes better than the categorized LDL, since all of its fittings below 100 mg/dL are consistently over estimated, except for the outlier. Between 100 and 190 mg/dL, the linear, log, quadratic, and spline model behave similarly. The categorical model deviates the most from those most centered LDL values. After 190 mg/dL the spline model and the quadratic model behave similarly.**
		2. 
		3. 
	3. *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?
		1. **For the sake of precision, an a priori method would have been the Cox proportional hazards regression model since then I would be able to model mortality over all time of the study. Furthermore, I would have preferred the spline method, mainly because the Mayo clinic guidelines do exist and so we should already have reason to believe that those prescribed groups require controlling over. Also, while it possible that there is heteroskedasticity within the relationship between mortality and LDL, going with the spline approach may prove helpful in capturing that within group variance.**

**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.