**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.

**Ans**: There are several possible issues with performing this analysis. For routine logistic regression, the outcome needs to be binary, which in this case it is. The predictor can be continuous or categorical, however interpreting the slope of a model for increasing levels of "ethnicity" is scientifically meaningless. Therefore, I chose to perform an analysis that generates odds ratios and inference for each group of ethnicity in reference to white subjects. I performed logistic regression with robust methods allowing for unequal variances. Statistical inference was based on the Wald statistic computed from the regression slope parameter and its standard error, with two-sided p value and 95% confidence interval computed using the approximate normal distribution for logistic regression parameter estimates. I then performed chi2 test to assess for significance of the overall model.

This is a saturated model, in that there are equal numbers of parameters reported (4) as there are groups.

.

Of the 735 subjects analyzed, 572 were white, 104 were black, 47 were Asian and 12 had race listed as other. The proportions of each racial group who had diabetes at the time of enrollment were: 9.79% of whites, 17.3% of blacks, 6.38% of Asians, and 16.7% of "other". Testing the significance of the overall model produces a p-value of 0.0956. This suggests that overall, there is no association between race and diabetes.

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

**Ans:** The odds ratios reported are the exponentiation of the slope (ie logOR) that was generated by the logistic model. The intercept reported is the odds of diabetes in the referent group, in this case whites, exponentiated from the log odds produced in the model. It agrees with the odds as calculated from the probability of diabetes in whites (0.0979/1-0.0979).

In comparison to white subjects, the OR for blacks was 1.93, with p value 0.026 and 95% CI [1.08, 3.44]. This observed OR for diabetes would not be unusual if the true odds for diabetes were between 1.08 and 3.44. In comparison to whites, the OR for Asian subjects was 0.628, with p value 0.449 and 95% CI [0.189, 2.09]. This observed OR would not be unusual if the true odds for diabetes were between 0.189 and 2.09. However, this was not statistically significant, with p value >0.05 and 95% CI for the OR that crosses 1. In comparison to whites, the OR for "other" ethnicity was 1.84, with p value 0.438 and 95% CI [0.393, 9.63]. Again, the wide CI and high p-value show that in comparison to whites, rates of diabetes is "other" ethnicity patients was not statististically significant. We therefore reject the null hypothesis that the odds of diabetes in whites and blacks are the same. We cannot reject the null hypothesis that the odds of diabetes are the same in other ethnic groups.

* 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.

**Ans**: I would conclude that while the overall model does not show significance, some of the individual comparison groups are significant. Only the odds ratio of whites versus blacks is statistically significant with p<0.05.

* 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)?

**Ans:** I recoded the variable for ethnicity such that race of black was now the referent group, then ran logistic regression with robust methods allowing for unequal variances. Again, I chose to perform an analysis that generates odds ratios and inference for each group of ethnicity in reference to black subjects. Statistical inference was based on the Wald statistic computed from the regression slope parameter and its standard error, with two-sided p value and 95% confidence interval computed using the approximate normal distribution for logistic regression parameter estimates. I then performed chi2 test to assess for significance of the overall model.

Of the 735 subjects analyzed, 572 were white, 104 were black, 47 were Asian and 12 had race listed as other. The proportions of each racial group who had diabetes at the time of enrollment were: 9.79% of whites, 17.3% of blacks, 6.38% of Asians, and 16.7% of "other". Testing the significance of the overall model produces a p-value of 0.0956. This suggests that overall, there is no association between race and diabetes.

This model is a reparamaterization of the model in part a. However, the p-value generated is the same and therefore the statistical inference for the overall model is the same.

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

**Ans:** The intercept represents the odds of diabetes in black subjects, and is the exponentiation of the intercept generated by the logistic regression model. The ORs reported are the exponentiation of the slopes of the logistic regression model.

The value of OR given for whites in reference to blacks has the same p value for inference, and is equal to the reciprocal of the OR given in part a. The OR of whites compared to blacks is 0.519, with p value 0.026 and 95% CI [0.291, 0.924].

However, the other odds ratios given represent new comparisons not performed in the analysis in part a. The OR of Asians in reference to blacks is 0.326, with p value 0.085 and 95% CI [0.091, 1.166]. This suggests that the observed OR is not unusual if the true OR lies between 0.091 and 1.166. The OR of "other" races in reference to Blacks is 0.956, with p value 0.956 and 95% CI [0.192, 4.74]. This suggests that the observed OR is not unusual if the true OR lies between 0.192 and 4.74. In this case, we would reject the null hypothesis that there is no difference in odds of diabetes between whites and blacks, with whites having a lower odds of diabetes. We would fail to reject the null hypothesis that there is a difference in odds of diabetes between blacks and asians or those of other races.

* 1. If we were to ignore issues 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.

**Ans**: I would conclude that while the overall model does not show significance, some of the individual comparison groups are significant. Only the odds ratio of whites versus blacks is statistically significant with p<0.05, with same p value as in part a but reciprocal odds.

* 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)?

**Ans:** Although these two models are just reparamaterizations of each other, and provide the same p-value for the overall model, the p-values for each of the odds ratios generated are different between the two models. The fact that we generate different p values depending on who is the referent group suggests that this would be an unreliable method to determine the significance of variables and should not be used alone to make decisions regarding dropping variables.

.

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.



|  |  |
| --- | --- |
|   | Proportion Alive |
| LDL level, mg/dl | 2 years | 5 years |
| LDL <70 | 100% (referent) | 0.591 [0.361,0.762] |
| LDL 70-100 | 0.958 [0.909, 0.981] | 0.832 [0.760, 0.884] |
| LDL 100-130 | 0.939 [0.899, 0.963] | 0.811 [0.754, 0.857] |
| LDL 130-160 | 0.956 [0.919, 0.976] | 0.871 [0.819, 0.909] |
| LDL 160-190 | 0.988 [0.9175, 0.9983] | 0.879 [0.788, 0.933] |
| LDL>190 | 0.958 [0.739, 0.994] | 0.833 [0.615, 0.934] |

.

**Ans**: I created a categorical variable representing LDL level cut at the points described by the Mayo Clinic: LDL <70, 70-99, 100-129, 130-159, 160-189, and >=190 mg/dl. I performed cox proportional hazards regression analysis on these groups of subjects, with robust methods, excluding those subjects without LDL measurements. The subjects were first analyzed by kaplan meier methods, examining observation time and death versus censoring. LDL level was then used as a categorical predictor, assessing the instantaneous hazard of death throughout the observation time, each in reference to the LDL<70 group. Statistical inference on the difference in probabilities of death was based on the Wald statistic computed from the regression slope parameter and its standard error as estimated using the Huber-White sandwich estimator, with two-sided p value and 95% confidence interval computed using the approximate normal distribution for linear regression parameter estimates. These results were then analyzed by chi2 methods to determine significance of the overall model.

725 subjects were observed, as 10 patients without LDL values were excluded. See table above for proportions alive at two years and for CIs. The proportion alive at five years in each category of LDL was 0.590 in those with LDL<70, 0.832 in those with LDL 70-99, 0.811 in those with LDL 100-129, 0.871 in those with LDL 130-159, 0.879 in those with LDL 160-189, and 0.833 in those with LDL>190. Testing the significance of the overall model produced a chi2 p value of 0.0191, suggesting the overall model is significant, and there is an association with LDL level and death.

3/3 for descriptive statistics

3/3 for performing an appropriate analysis

1/4 for reporting the association appropriately

Wrong interpretation of coefficient (-2)

Wrong interpretation of CI (-1)

Total: 7

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

**Ans**: See above for the values of the various hazards. The hazards calculated by the model represent the exponentiation of the slope of the hazards model. Typically, no intercept is reported in proportional hazards analysis. However, if an intercept were reported, it would refer to the "baseline" hazards, in this case in the referent group of subjects with LDL = 0

Each slope is the hazard ratio referent to baseline hazard for LDL 0 to 69. The hazard ratio of the patients with LDL 70-99 was 0.398, with p value 0.008 and 95% CI [0.203, 0.782]. The HR for patients with LDL 100-129 was 0.393 with p value 0.004 and 95% CI [0.207, 0.744]. The HR for patients with LDL 130-159 was 0.293, with p value <0.001 and 95% CI [0.152, 0.568]. The HR for subjects with LDL 160-190 was 0.257, with P value 0.001 and 95% CI [0.113, 0.579]. The HR for subjects with LDL >190 was 0.317, with p value 0.048 and 95% CI [0.101, 0.989]. This suggests a progressively decreasing hazard ratio for death in comparison to the LDL<70 group with increasing strata of LDL level, all of which are statistically significant with pvalues<0.05. However, the CIs are wide.

This question is asking you to do interpretation. You just wrote down the numbers without telling what those number means.

Total: 0

* 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?

**Ans**: We could compare the overall R2 in the two models to see which one explains more of the variation. Another method would be to perform the regression analysis using the continuous linear term as a predictor variable, then also include the categorical dummy variables, and then use the results to perform a chi2 test for significance of nonlinearity on the output of that model.

Performing this analysis, though it does drop one category of the dummy variables, yields a p-value of 0.470. This would not indicate a better fit for the categorical variables than the linear continuous model.

Did not mention about the test that regression coefficients for the dummy variables were 0 (-1)
Did not mention what kind of test you use (-1)
Wrong p-value (-1)
Wrong conclusion (-1)

Total: 1

* 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.

**Ans**: I created a group of splines representing LDL level cut at the points described by the Mayo Clinic: LDL <70, 70-99, 100-129, 130-159, 160-189, and >=190 mg/dl. I performed cox proportional hazards regression analysis on these groups of subjects, with robust methods, excluding those subjects without LDL measurements. The subjects were first analyzed by kaplan meier methods, examining observation time and death versus censoring. The splines of LDL were then used as predictors, assessing the instantaneous hazard of death throughout the observation time. Statistical inference on the difference in probabilities of death was based on the Wald statistic computed from the regression slope parameter and its standard error as estimated using the Huber-White sandwich estimator, with two-sided p value and 95% confidence interval computed using the approximate normal distribution for linear regression parameter estimates. These results were then analyzed by chi2 methods to determine significance of the overall model.

725 subjects were observed, as 10 patients without LDL values were excluded. See table above for proportions alive at two years and for CIs. The proportion alive at five years in each category of LDL was 0.590 in those with LDL<70, 0.832 in those with LDL 70-99, 0.811 in those with LDL 100-129, 0.871 in those with LDL 130-159, 0.879 in those with LDL 160-189, and 0.833 in those with LDL>190. The overall model was significant, with chi2 p value <0.0001, supporting an association between LDL level and death.

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

**Ans**: Again, there is not really an intercept reported in proportional hazards, but the intercept would represent the hazard in the baseline group. The slopes reported represent the hazard ratio between two groups within the defined knots, with differences of one unit of LDL.

Therefore, the hazard ratio modeled within the splines for LDL <70 is 0.978, with p value 0.019 and 95% CI [0.960, 0.996], for subjects with LDL 70-99 was 0.979, with p value 0.139 and 95% CI [0.953, 1.01]. The HR for patients with LDL 100-129 was 0.998 with p value 0.835 and 95% CI [0.0.976, 1.019]. The HR for patients with The HR for patients with LDL 130-159 was 1.004, with p value <0.773 and 95% CI [0.979, 1.028]. The HR for subjects with LDL 160-189 was 0.971, with P value 0.181 and 95% CI [0.929, 1.013]. The HR for subjects with LDL >=190 was 1.029, with p value 0.261 and 95% CI [0.979, 1.081].

These data are interesting, in that the p values for the individual hazards within the knots of the splines do not indicate significance, but testing of the overall model does indicate significance.

* 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?

**Ans:** Similar to above, we could perform the proportional hazards regression analysis using the continuous linear term for LDL as a predictor variable, then also include the splines as predictors, and then use the results from the spline output to perform a chi2 test for significance of nonlinearity.

Performing this analysis, though it does drop one category of the dummy splines, it yields a p-value of 0.0788. This would not indicate a that the spline model is better fit than the continuous model and we do not have evidence for nonlinearity.

* 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?

**Ans**: The methods used in HW 1-3 depended primarily on dichotomizing the data at a set cutpoint (160 mg/dl of LDL). Changing this continuous variable into a dichotomous variable causes it to lose a great deal of precision, and potentially to mask associations that are in fact present due to this lack of precision.

The methods used in HW 4 and 5 better preserve of the continuous nature of the data (preserving power by preserving DOF). Using K-M methods allows the survival to be examined throughout the full study period, not just until the first censoring event. HW 5 uses "flexible" methods that would allow for wider variety of non-linear relationships in the data, which can helpful if a non-linear relationship is suspected.

* 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.

**Ans**: The spline model is similar to the quadratic and log models from HW 4, in that it has a ssignificant upward trend at the extremes in data where the number of subjects is fewer. The Categorical model retains somewhat of a this u-shape trend, but is particularly exaggerated in the low values of LDL. The spline model demonstrates an upward trend around 150 mg/dl, which is not found on any of the other models. This reflects the flexible nature of this model.



* 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?

**Ans:** I would prefer the methods used in HW 4, using LDL as a log-transformed predictor and using K-M methods with proportional hazards regression. Overall, that model keeps the continuous nature of the data, and provides a more linear relationship. I would use cox proportional hazards to utilize as much of the censored data as possible, without any arbitrary cut points. This method provides inference on an easily interpretable scale. While the methods presented in this HW are flexible, they provide output that is very difficult to interpret.

**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.