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.

This is a saturated model since we have two data points (white or non-white) and the same parameters for estimation.

Method: Distributions of prevalence of diabetes from race was compared across white race group and non-white race group using logistic regression modeling and the indicator of black, asian and other as the dummy variable. Quantification of association between race group computed from the regression model, with confidence intervals and two-sided p values computed.

Inference: Data were available on 735 subjects, in which we have 572 white and 163 non-white subjects. In total, we have 79 people have diabetes. Among them, 56, 18, 3, 2 are from white, black, asian and other race group. From a logistic regression analysis, we estimate that the slope for black, asian and white are .6146, .3409, 1 respectively and intercept is 0.2

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

The slope indicates that the odds ratio between prevalence of diabetes in black, asian and other people and white people is 1.929 (CI: 1.082-3.439), 0.6282 (CI: 0.1888-2.081), 1 respectively, relative. The intercept, 0.1085, is the prevalence of diabetes for the white people (CI: 0.8236, 0.1430). All the explanation is the same for all those slope and CI. Here I only show one example: slope for black is the odd ratio between black race group and white race group. black group will have black equal to one and asian, other equal to zero while white group will have black, asian and other as zero. Based on a 95% confidence interval, the slope suggesting the odds ratio of prevalence of diabetes between black people and white people, 0.6146 would not be judged unusual if the true prevalence of diabetes were anywhere from 1.082 to 3.439. A two-sided p value of 0.026 suggests that we have high confidence reject the null hypothesis that the odds ratio is from 1. Note that other group doesn’t such a high confidence to reject the null hypothesis that the odds ratio is from 1.

* 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.

The odds ratio of diabetes prevalence between asian and white subjects are not significantly different from 1 while the two side p-value reported here are far more than 0.05. However, that of the black people has significant p-value (0.026) which indicates the high association between white people and diabetes prevalence.

* 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)?

This time we are using the black group rather than the white group as the reference group. Since the underlying model is the same although we choose different dummy variables, the overall test of association between diabetes and race group are the same with the chi-square p-value 0.0956. the odds ratio, CI and the p-value between the prevalence of diabetes in group asian, other to the reference group are changed in two models between of the choice of different reference group while the p-value (0.026) are hold the same for the white or black to the reference group and the odds ratio and the corresponding CI are reciprocal inverse between two models.

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

The slope indicates that the odds ratio between prevalence of diabetes in asian, other and white people and black people is 0.3258 (CI: 0.09094-1.167), 0.9556 (CI: 0.1925-4.742), 1 respectively, relative. The intercept, 0.2093, is the prevalence of diabetes for the black people (CI: 0.1259, 0.3480). All the explanation is the same for all those slope and CI. Here I only show one example: slope for white is the odd ratio between white race group and black race group. white group will have white equal to one and asian, other equal to zero while black group will have white, asian and other as zero. Based on a 95% confidence interval, the slope suggesting the odds ratio of prevalence of diabetes between black people and white people, 0.5185 would not be judged unusual if the true prevalence of diabetes were anywhere from 0.2908 to 0.9246. A two-sided p value of 0.026 suggests that we have high confidence reject the null hypothesis that the odds ratio is from 1. Note that other group doesn’t have significant p-value.

* 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.

The odds ratio of diabetes prevalence between other, asian subjects are not significantly different from 1 while the two side p-value reported here are far more than 0.05.However, that of the white people has significant p-value (0.026) which indicates the high association between white people and diabetes prevalence.

* 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)?

The p-value generated from different model, like the different models with distinct reference as shown here, will have different values (excepting the corresponding group regarding to the switch of the reference group). If we are going to drop certain variables, it may lead to lose information and end up with wrong regression. We may need to use Root Mean Square Error (RMSE) to decide the validity of our model

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

|  |  |  |
| --- | --- | --- |
|  | Serum LDL descriptive statistics  | All Subjects (with LDL Available3) |
|  | 11 – 69 mg/dL | 70 – 99 mg/dL | 100 – 129 mg/dL | 130 – 159 mg/dL | 160 – 189 mg/dL | 190 – 247 mg/dL |
| N Subjects | 22 | 143 | 228 | 225 | 83 | 24 | 725 |
| N Deaths | 10 | 28 | 44 | 34 | 11 | 4 | 131 |
| 2 year Survival Probability | 100% | 95.8% | 93.9% | 95.6% | 98.8% | 95.8% | 95.6% |
| 5 Year Survival Probability | 59.1% | 83.2% | 81.1% | 87.1% | 88.0% | 83.3% | 83.6% |
| 10th Pctile of Survival  | 3.46 y | 3.80 y | 3.41 y | 4.30 y | 4.53 y | 4.13 y | 3.66 y |
| 20th percentile of Survival | 3.55 y | 5.44 y | 5.36 y | NA1 | NA1 | NA1 | 5.54 y |
| 5.75 Year Restricted Mean of Survival \*2 | 4.91 y | 5.24 y | 5.23 y | 5.35 y | 5.45 y | 5.32 y | 5.29 y |

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.

In table 1 we estimate the survival distribution based on the Mayo clinic categories. From the total 725 subjects, 133 deaths are observed during the study process. The figure 1 presents the Kaplan-Meier estimates for different ldl groups defined by the Mayo Clinic.

Method: Distributions of hazard of death over the entire period of observation across groups defined by serum LDL was compared across different groups as defined by Mayo Clinic LDL categories, using hazard ratio regression modeling and the categories as defined by the Mayo

clinic. Quantification of association between race group computed from the regression model, with confidence intervals and two-sided p values computed.

**Table 1. descriptive statistics for the people in study**

**Figure 1 Kaplan-Meier estimates for the ldl group**

Inference: Data were available on 735 subjects with mean serum LDL as 126 mg/dL (SD 33.6 mg/dL, min: 11 mg/dL and max: 247 mg/dL). 131 subjects are dead during an average of 5.33 years. The total p-value from the hazard ratio regression is 0.0087 and thus we can reject the null hypothesis that there is not any association between the LDL and death.

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

We used the regression to calculate the hazards ratio in different group. ldlCTG is used as the dummy variable. The hazard ratio for people with ldl mg/dL, between 70 mg/dL and 100 mg/dL, between 100 mg/dL and 130 mg/dL, between 130 mg/dL and 160 mg/dL and between 160 mg/dL and 190 mg/dL are 0.3980, 0.3926, 0.2939, 0.2565, 0.3167, respectively. All the p-value for each category is less than 0.05 which implies the statistically significant difference from 1 for the hazard ratio in each group comparing the baseline group for the ldl from 0 to 69mg/dL. Since all the explanation for the CI are the same for each group, I will also talk about the group 70. Based on a 95% confidence interval, the slope suggesting the hazard ratio between group 70 people and the baseline group, 0.3980. would not be judged unusual if the true prevalence of diabetes were anywhere from 0.2026 to 0.7820. A two-sided p value of 0.008 suggests that we have high confidence reject the null hypothesis that the hazard ratio is from 1. Note that overall association between the ldl and death from the cox regression has significant p-value as 0.0087.

* 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?

RMSE (root mean square error) should be used to assess this question. for a better fit model, the RMSe should be much less comparing this ldl-discreted regression model with dummy variable. From the comparison of the RMSE, we can conclude that the linear model with ldl as continuous term is a better model.

* 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. the result shows that the linear model with a continuos ldl has much less rmse and therefore should be a better fit model.
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.

See answer 2a for the descriptive statistics.

Method: Distributions of hazard of death over the entire period of observation across groups defined by serum LDL was compared across different groups as defined by Mayo Clinic LDL categories, using hazard ratio regression and the linear splines as indicated the question. Quantification of association between race group computed from the regression model, with confidence intervals and two-sided p values computed.

Inference: Data were available on 735 subjects with mean serum LDL as 126 mg/dL (SD 33.6 mg/dL, min: 11 mg/dL and max: 247 mg/dL). 131 subjects are dead during an average of 5.33 years. The total p-value from the hazard ratio regression is 0.0285 and thus we can reject the null hypothesis that there is not any association between the LDL and death.

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

We used the regression to calculate the hazards ratio in different group based on the linear splines as presented in the question. The hazard ratio for people with ldl below 70 mg/dL and between 70 mg/dL and 100 mg/dL, between 100 mg/dL and 130 mg/dL, between 130 mg/dL and 160 mg/dL and between 160 mg/dL and 190 mg/dL are 0.980, 0.998, 1.0036, 0.971, 1.029, respectively. All the p-value for each category is much higher than 0.05 which implies the statistically significant difference from 1 for the hazard ratio in each group comparing the baseline group for the ldl from 0 to 69mg/dL. Since all the explanation for the CI are the same for each group, I will also talk about the group 70. Based on a 95% confidence interval, the slope suggesting the hazard ratio between group 70 people and the baseline group, 0.3980. would not be judged unusual if the true prevalence of diabetes were anywhere from 0.9491 and 0.9967. A two-sided p value of 0.0139 suggests that we have high confidence reject the null hypothesis that the hazard ratio is from 1. Note that overall association between the ldl and death from the cox regression has significant p-value as 0.0087.

* 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?

RMSE (root mean square error) should be used to assess this question. for a better fit model, the RMSe should be much less comparing this ldl-discreted regression model with dummy variable.

From the comparison of the RMSE, we can conclude that the linear model with ldl as continuous term is a better mode.

* 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?

since we don’t dichotomize the data, we don’t lose information by the process of dichotomization. As a result, it allow us to do better job in assessing the association and much more power to predict the situations.

* 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.

In the following diagram, I demonstrate the hazard ratio obtained from different approaches used across homework 4 and this homework. It shows that the trend for the association is identical with a few outliers near the two sides.



* 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?

I will definitively choose the proportional hazard regression sussing the hazards ratio without any transformation to assess the linear association. I will then continue to address the non-linear association if I cannot show significant linear association in the proportional hazard regression analysis. the dummy variable and linear spline regression can give me deeper understanding on those non-linear trends.

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.