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.

It is a saturated model, since we modeled four groups (four races) with three predictors plus intercept.

Methods: The odds of a diabetes diagnosis were compared between groups of subjects defined by race using a logistic regression model that uses whites as a reference group. Statistical inference was based on confidence intervals and two-sided p-values that were computed using Wald statistics based on the Huber-White sandwich estimator.

Inference: A two-sided p-value of 0.0956 suggests that we cannot reject the null hypothesis that odds of having diabetes are not associated with race.

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

The intercept is the odds of having diabetes for subjects with race “white” and the slopes are the odds ratios of having diabetes for the groups of subjects with races “black”, “asian”, and “other” compared to the group of subjects with race “white”.

* 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 subjects in the group with race “white” are estimated to have the odds of a diabetes diagnosis of 0.109 (95% CI unadjusted for multiple comparisons: 0.082 to 0.143). Compared to the subjects in the group with race “white”, the odds ratio for the group of subjects with race “black” was found to be 1.929 (95% CI unadjusted for multiple comparisons: 1.082 to 3.439), the odds ratio for the group of subject with race “asian” was 0.628 (95% CI unadjusted for multiple comparisons: 0.189 to 2.091) and the odds ratio for the group of subjects with race “other” was 1.843 (95% CI unadjusted for multiple comparisons: 0.393 to 8.631). Two-sided p-values of 0.449 and 0.438 for “asian” and “other” groups respectively suggest that we cannot reject the null that the odds of a diabetes diagnosis are different for the subjects in the groups with race “asian” and race “other” compared to the subjects in the group with race “white”. However, a two-sided p-value of 0.026 for the groups of subjects with race “black” suggests that the odds of diabetes are higher in subjects with race “black” than subjects with race “white”.

* 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 report of formal inference would be essentially the same (overall p-value would be the same), since this model is a reparameterization of a logistic regression model that uses whites as a reference group.

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

The intercept is the odds of having diabetes for subjects with race “black” and the slopes are the odds ratios of having diabetes for the groups of subjects with races “white”, “asian”, and “other” compared to the group of subjects with race “black”.

* 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 subjects in the group with race “black” are estimated to have the odds of a diabetes diagnosis of 0.209 (95% CI unadjusted for multiple comparisons: 0.126 to 0.348). Compared to the subjects in the group with race “black”, the odds ratio for the group of subjects with race “white” was found to be 0.519 (95% CI unadjusted for multiple comparisons: 0.291 to 0.925), the odds ratio for the group of subject with race “asian” was 0.326 (95% CI unadjusted for multiple comparisons: 0.091 to 1.167) and the odds ratio for the group of subjects with race “other” was 0.956 (95% CI unadjusted for multiple comparisons: 0.193 to 4.742). Two-sided p-values of 0.085 and 0.956 for “asian” and “other” groups, respectively, suggest that we cannot reject the null that the odds of a diabetes diagnosis are different for the subjects in the groups with race “asian” and race “other” compared to the subjects in the group with race “black”. However, a two-sided p-value of 0.026 for the groups of subjects with race “white” suggests that the odds of diabetes are higher in subjects with race “white” than subjects with race “black”.

* 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)?

If we only base our decision on one of the models, we might incorrect in our decision to include or exclude certain variables in a regression model. The first model only gives us three out of six pairwise comparisons. Therefore, if one of the variables is statistically insignificant, it might be significant in the other 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

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:

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.

Descriptive statistics: The study included 735 subjects who were followed for death from any cause. Serum LDL measurements were not available on 10 subjects, and therefore were excluded from the analyses. In the 725 subjects with available serum LDL measurements at enrollment, the mean LDL was 126mg/dL (SD 33.6 mg/dL, range 11 to 247 mg/dL).

Table 1 presents estimates of the survival distribution within groups defined by serum LDL. The biggest difference in survival probabilities is between the group with the lowest serum LDL after 2 years of follow-up. The 5-year survival probability is the lowest for that group at 59.1% and the highest for the group with serum LDL between 160 and 189mg/dL (88.0%).

Figure 1 shows the Kaplan-Meier survival probability estimates for strata defined by serum LDL. From the graph, we can see that the group with the lowest LDL has the survival distribution different from other groups.

Table 1. Kaplan Meier based estimates for subjects having serum LDL measurements at baseline.

|  |  |
| --- | --- |
|  | Serum LDL at enrollment  |
|  | 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 |
| N deaths | 10 | 28 | 44 | 34 | 11 | 4 |
| 2yr Survival Probability | 1.00 | .958 | .939 | .956 | .988 | .958 |
| 5yr Survival Probability | .591 | .832 | .811 | .871 | .880 | .833 |

Figure 1: Kaplan-Meier based estimates of distribution of time from study enrollment to death from any cause for 725 subjects having serum LDL measurements at baseline.



Methods for inferential statistics: The instantaneous risk (hazard) of death over the entire period of observation across groups defined by serum LDL were compared using Cox proportional hazards with the groups fit as dummy variables. Statistical inference was based on p-values and 95% CIs that were computed using Wald statistics based on the Huber-White sandwich estimator. Subjects missing data on serum LDL were omitted from the analysis.

Results: A two-sided p-value of 0.009 suggests that we can with high confidence reject the null hypothesis that the risk of death is not associated with serum LDL levels.

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

The intercept is a baseline hazard, or in our case it is an estimated instantaneous risk of death over the entire period of observation for the group with serum LDL between 0 and 69mg/dL. It is not given in the output in STATA but can be calculated.

The slope is a hazard ratio for a group compared to the baseline group. Compared to the hazard of death for the group with serum LDL between 0 and 69mg/dL, we estimate that the group with serum LDL between 70mg/dL and 99mg/dL has the instantaneous risk of death of 60.2% lower, the group with serum LDL between 100mg/dL and 129mg/dL has the hazard of death 60.7l% lower, the group with serum LDL between 130 and 159mg/dL has the hazard of death 70.6% lower, the group with serum LDL between 160 and 189mg/dL has the hazard of death 74.3% lower, and the group with serum LDL greater than or equal to 190 has the hazard of death 68.3% lower.

* 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?

We can use a chi-square test to see if there is any evidence for nonlinearity when using dummy variables to detect it. A two-sided p-value of 0.3988 suggests that there is not enough 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. 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.

For descriptive statistics, refer to question 2a.

Methods for inferential statistics: The instantaneous risk (hazard) of death over the entire period of observation across groups defined by serum LDL were compared using Cox proportional hazards with the groups fit as linear splines. Statistical inference was based on p-values and 95% CIs that were computed using Wald statistics based on the Huber-White sandwich estimator. Subjects missing data on serum LDL were omitted from the analysis.

Results: A two-sided p-value < 0.001 suggests that we can with high confidence reject the null hypothesis that the risk of death is not associated with serum LDL levels.

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

The intercept is an estimated instantaneous risk of death for subjects with LDL equal to 0. The estimate does not have any scientific meaning.

The slopes are estimated differences in the instantaneous risks of death between two groups both between the same knots but differing in serum LDL by 1 mg/dL.

From Cox proportional hazard model, we estimate the group with the lowest LDL to have the hazard of death to be lower by 2.19% for every increase in serum LDL by 1mg/dL. For the group with serum LDL between 70 and 99mg/dL, we estimate the hazard of death to be lower by 2.03% for every 1mg/dL increase in serum LDL. Similarly, for the group with serum LDL between 100 and 129mg/dL, we estimate the hazard of death to be lower by 0.23% for every 1mg/dL increase in serum LDL. For the subjects with serum LDL between 130 and 159mg/dL, the instantaneous risk of death is estimated to be higher by 0.36% for every 1mg/dL increase in serum LDL. . For the group with serum LDL between 160 and 189mg/dL, we estimate the hazard of death to be lower by 2.91% for every 1mg/dL increase in serum LDL. Lastly, for the subjects with serum greater than or equal to 190mg/dL, the instantaneous risk of death is estimated to be higher by 2.88% for every 1mg/dL increase in serum LDL.

* 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?

We can use a chi-square test to see if there is any evidence for nonlinearity when using splines. A two-sided p-value of 0.0788 suggests that there is not enough 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?

Stratifying data allows us to see patterns within smaller groups and can potentially adjust for confounding. They are also advantageous is there is suspicion of a nonlinear fit.

* 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.

Figure 2 displays the fitted values from five models. In each case, there is a downward trend with higher LDL. The greatest difference between models occur with the lowest values of LDL while all five models are similar over the midrange of LDL.

Figure 2. Plots of the fitted hazard ratios from 5 models.



* 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 would prefer proportional hazards regression with log link. Using log transformed data allows for easy interpretation of hazard ratios. It also addresses the question of nonlinearity. However, it is not as complex as dummy variables or splines and we do not lose any precision by performing a univariate transformation.

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.