**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.
* Yes. It’s a saturated model because we have total 4 parameters (3 parameters + intercept) after dummy and have 4 groups for race.
* Seeing that the outcome (diabetes) is a binary variable and we have 4 groups of the race, we can use logistic regression model with robust standard error to see the association between diabetes and race. Before the modeling, we should dummy the variable of race with “white“ as the reference. Thus we have 3 parameters to capture the race. The results showed the odds ratio and two-sided p-value and 95%CI.
* From the results, we know the odd ratio is 1.9286, 95%CI: (1.0815, 3.4391), p-value=0.026 when compare race=black to race=white; OR=0.6282, 95%CI=(0.1888, 2.0909), p-value=0.449 when compare race=Asian to race=white; OR=1.8429, 95%CI=(0.3935, 8.631), p-value=0.438 when compare race=Other to race=white. That is, race=black is 1.9286 times more likely to have diabetes than race=white. With 95% confidence, It is not unusual if the true OR is between 1.0815 and 3.4391 when compare black to white. For Asian, they are 37.2% lower to have diabetes than white, and 95% suggested it is not surprised if the true OR between 0.1888 and 2.0909 when compare Asian and white. For other race, they are 1.8429 times more likely to have diabetes than white. With 95% confidence, we will not be surprised if the true OR is between 0.3935 and 8,631 when compare other race and white. The two-sided p-value for intercept <0.0001 and as mentioned above, we know the p-value for each dummy variable. Seeing that not all of the p-value are statistically significant under the alpha level=0.05. We have no evidence to show the 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).
* After dummy, our model will become:

$logodd(diabetes|race)=β\_{0}+β\_{1}race1+β\_{2}race2+β\_{3}race3$.

 Then, race1=race2=race3=0 when race= white

 race1=1, race2=race3=0 when race=black

 race2=1, race1=race3=0 when race=Asian

 race3=1, race1=race2=0 when race=other

Therefore,

* $β\_{0}=intercept.$ It means the log odd of diabetes when race=white because the reference group is white; the $β\_{1}$ means the different of log odd of diabetes when race=black compare to white, so $e^{β\_{1}}=odds ratio.$ The $β\_{2}$ means the different of log odd of diabetes when race=Asian compare to white; the $β\_{3}$ means the different of log odd of diabetes when race=other compare to 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.
* From part a. the p-value for race=Asian is 0.449 and for race=other is 0.438 are not statistically significant under the alpha=0.05. If we ignore the multiple comparisons, we can say the odd of diabetes is difference between black and white. Black tends to have higher risk to have diabetes because the odds ratio is 1.9286 and two-sided p-value=0.026 < alpha=0.05. And when building model, we can keep the variables with statistically significant in the model.
	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)?
* When change the reference group to from black, we can get the OR= 0.5185 when compared white to black with 95%CI: (0.2908, 0.9246), p-value=0.026; OR=0.3258 with 95% CI: (0.0909, 1.1669), p-value=0.085 when compared Asian to black; OR=0.9556 with 95%CI (0.1925, 4.7324),p-value=0.956 when compared other to black. We know white is 48.15% less likely to have diabetes than black, and under 95% confidence, it is not surprised if the true OR between 0.2908 and 0.9246. When compared Asian to black, Asian is 67.42% less likely to have diabetes, and we are not surprised if the true OR is between 0.0909 and 1.1669 with 95% confidence. For other race, we know say other race is 4.44% less likely to have diabetes than black. It is not unusual if the true OR is between 0.1925 and 4.7324. The p-value for intercept < 0.0001, but not all of the two-sided p-value for each dummy variable are statistically significant. There is no evidence to say the association between race and diabetes.
* The reciprocal of OR=0.5185 that is compare black to white exactly agree with the OR=1.9286 when compare white to black.
	1. Using the regression model fit in part (d), provide an interpretation for each of the regression parameters (including the intercept.)
* When changed the reference group to black, we know:

 race1=race2=race3=0 when race= black

 race1=1, race2=race3=0 when race=white

 race2=1, race1=race3=0 when race=Asian

 race3=1, race1=race2=0 when race=other

* Unlike part a, the intercept means the log odd of diabetes when race=white because the reference group is black here; the $β\_{1}$ means the different of log odd of diabetes when race=white compare to black, so $e^{β\_{1}}=odds ratio.$ For $β\_{2}$, it means the different of log odd of diabetes when Asian compared to black and the $β\_{3}$ means the different of log odd of diabetes when other race compared to 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.
* If we ignore the multiple comparisons, we can say white is less likely to have diabetes than black because of its OR=0.5185 and its two-sided p-value=0.026 < alpha=0.05, having statistically significant. And when building model, we can keep the variables with statistically significant in the model.
	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 do that, the type 1 error will increase.
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.
* Because the interested outcome is hazard of death, we can use cox regression model. The LDL is a continuous variable, but we will dummy it by following the Mayo clinic’s cut point, to divide it to six groups. Then we can get the hazard ratio for each group compare to the reference group, and the two-sided p-value and 95% CI.
* We know the sample size is 725 and most of them have LDL between 100-129 mg/dL (31.45%) or 130-159 mg/dL (31.25%). We can see the distribution of LDL between deaths and censored in the following table. There are total 131 people died in this dataset, and 594 people censored.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Censored (N=594) | Death N=131 | Total N=725 |
| LDL | n (%) | n (%) | n (%) |
| Below 70 mg/dL | 12 (2.02) | 10 (7.63) | 22 (3.03) |
| 70-99 mg/dL | 115 (19.36) | 28 (21.37) | 143 (19.72) |
| 100-129 mg/dL | 184 (30.98) | 44 (33.59) | 228 (31.45) |
| 130-159 mg/dL 130 | 191 (32.15) | 34 (25.95) | 225 (31.03) |
| 160-189 mg/dL 160 | 72 (12.12) | 11 (8.4) | 83 (11.45) |
| Higher than 190 mg/dL | 20 (3.37) | 4 (3.05) | 24 (3.31) |

* By using Kaplan Meier method to calculate the survival function for year 1, year 3, and year 5 , we can see the survival is lower when people’s LDL below 70 mg/dL at 5 year, the estimator of survival function is 0.59509 with 95% CI: (0.361, 0.7621). And there is no too much difference between other groups.



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | S(t) | 95% CI |  | S(t) | 95% CI |
| below 70 mg/dL  |  |  | 130- 159 mg/dL  |  |  |
| 1 | 1.0000 |  | 1 | 0.9778 | (0.9474, 0.9907) |
| 3 | 0.9091 | (0.683, 0.9765) | 3 | 0.9289 | (0.8865, 0.9558) |
| 5 | 0.5909 | (0.361, 0.7621) | 5 | 0.8711 |  |
| 70-99 mg/dL  |  |  | 160-189 mg/dL  |  |  |
| 1 | 0.979 | (0.9364, 0.9932) | 1 | 1.0000 |  |
| 3 | 0.9091 | (0.8486, 0.9462) | 3 | 0.9639 | (0.8921, 0.9882) |
| 5 | 0.8322 | (0.7601, 0.8842) | 5 | 0.8795 | (0.7876, 0.9333) |
| 100-129 mg/dL  |  |  | higher than 190 mg/dL  |  |
| 1 | 0.9825 | (0.9539, 0.9934) | 1 | 1.0000 |  |
| 3 | 0.9123 | (0.8673, 0.9425) | 3 | 0.9167 | (0.7061, 0.9785) |
| 5 | 0.8114 | (0.7543, 0.8565) | 5 | 0.8333 | (0.6148, 0.9339) |

We used the cox regression model to see the association between mortality and LDL. From the results, we know the hazard ratio is 0.3980, 0.3926, 0.2939, 0.2565, 0.3167 for subjects whose LDL between 70-99 mg/dL, between 100-129 mg/dL, between 130-159 mg/dL, between 160-189 mg/dL and higher than 190 mg/dL, respectively. That is, the risk of death for people who have LDL between 70-100 mg/dL is 60.2% lower than those LDL below 70 mg/dL; the risk of death for people who have LDL between 100-129 mg/dL is 60.7% lower than who have LDL below 70 mg/dL; for people’s LDL between 130-159 mg/dL, for people’s LDL between is 160-189 mg/dL, and for people’s LDL higher than 190 mg/dL, the risk of death for them is 70.61% lower, 74.35% lower and 68.33% lower than people whose LDL below 70 mg/dL, respectively. All of the p-value for dummy variables are smaller than alpha level=0.05, having statistically significant. We can say LDL is associated with mortality. With 95% confidence, we are not surprised if the true hazard of death is from 21.8% to 79.74% lower for people with LDL70-99 mg/dL than whose LDL below 70 mg/dL; it’s also not surprised if the true hazard of death is from 25.58% to 79.29% lower in the 100-129 mg/dL group than whose LDL below 70 mg/dL. And it’s unusual if true hazard of death is from 43.22% to 84.79% lower in the 130-159 mg/dL group, from 42.01% to 88.65% lower in the 160-189 mg/dL group, from 1.09% to 89.86% lower than whose LDL below 70 mg/dL.

3/3 for descriptive statistics

1/3 for performing an appropriate analysis

Did not report whether using Huber-White sandwich estimator or not (-1)

Did not report which statistic the statistical inference is based on (-1)

4/4 for reporting the association appropriately

Total: 8

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

$$λ=λe^{β\_{1}X\_{1}+β\_{2}X\_{2}+β\_{3}X\_{3}+β\_{4}X\_{4}+β\_{5}X\_{5}}$$

* When dummy LDL,

we let X1 =X2=X3=X4=X5 =0when LDL below 70 mg/dL

X1 =1, X2=X3=X4=X5=0 when LDL between 70-99 mg/dL

X2 =1, X1=X3=X4=X5=0 when LDL between 100-129 mg/dL

X3 =1, X1=X2=X4=X5=0 when LDL between 130-159 mg/dL

X4 =1, X1=X2=X3=X5=0 when LDL between 160-189 mg/dL

X5 =1, X1=X2=X3=X4=0 when LDL higher than 190 mg/dL

* Therefore,

$β\_{1}$ means the difference of log odds of death between group with LDL below 70 mg/dL and groups with LDL between 70-99 mg/dL.

$β\_{2}$ means the difference of log odds of death between group with LDL below 70 mg/dL and groups with LDL between 100-129 mg/dL.

$β\_{3}$ means the difference of log odds of death between group with LDL below 70 mg/dL and groups with LDL between 130-159 mg/dL.

$β\_{4}$ means the difference of log odds of death between group with LDL below 70 mg/dL and groups with LDL between 160-189 mg/dL.

$β\_{5}$ means the difference of log odds of death between group with LDL below 70 mg/dL and groups with LDL higher than 190 mg/dL.

The intercept means the odds of death for groups with LDL below 70 mg/dL.

log odds => log hazard (-2)

Total: 3

* 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 test if each dummy variable is equal to zero to see the linearity. P-value =0.0087, that means we can reject the null hypothesis. There are strong evidence to show that is nonlinear.

Did not mention including linear term (-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.
* When let LDL of 160 mg/dL as the reference, we can get the hazard ratio of groups with LDL below 70 mg/dL is 3.8984. That is, the odd of death for people with LDL below 70 mg/dL is 3.8984 times higher than group with 160 mg/dL, and the odd of death for people’s LDL between 70-99 mg/dL is 1.5517 times higher, for people’s LDL between 100-129 mg/dL is 1.5304 times higher, for people’s LDL between 130-159 mg/dL is 1.1458 times higher, and for whose LDL higher than 190 mg/DL is 1.2347 times higher than group with 160 mg/dL.
* From the results, we can know the group with 160 mg/dL has lowest odd of death among this six LDL levels.
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.
		+ We still use cox regression model, but use the different way to fit the predictor - LDL. Therefore, the descriptive statistics will be as same as the part a. in the Q2.
		+ For cox regression model, we get the hazard ratio = 0.9781 within the group with LDL below 70 mg/dL; hazard ratio=0.09797 within the group with LDL between 70-99 mg/dL; the hazard ratio=0.9977, 1.0035, 0.9709, and 1.0288 within the groups with 100-129 mg/dL, 130-159 mg/dL, 160-189 mg/dL and higher than 190 mg/dL, respectively. With 95% CI, we are not surprised if the true hazard ratio is between 0.9602 to 0.9963 within the group with LDL below 70 mg/dL; if true HR is between 0.9535 to 1.0067 within the group with LDL from 70-99 mg/dL; if true HR is between 0.9764 to 1.019 within the group with LDL between 100-129 mg/dL; if true HR is between 0.9794 to 1.0284 within the group with LDL between 130-159 mg/dL; if true HR is between 0.9298 to 1.01378 with the group with LDL from 160-189 mg/dL, and it’s not unusual if the true hazard ratio between 0.9791 to 1.081within the group with LDL higher than 190 mg/dL.
	2. Provide an interpretation for each parameter in your regression model, including the intercept.
* When fit model by splines, each parameter capture the difference within that group.

$$λ=λe^{β\_{1}LDL0+β\_{2}LDL70+β\_{3}LDL100+β\_{4}LDL130+β\_{5}LDL160+β\_{6}LDL190}$$

* Therefore,

$β\_{1}$ means the difference of log odds of death within the population with LDL below 70 mg/dL when their LDL changed by 1 units.

$β\_{2}$ means the difference of log odds of death within the groups with LDL between 70-99 mg/dL when their LDL changed by 1 units.

$β\_{3}$ means the difference of log odds of death within the groups with LDL between 100-129 mg/dL when their LDL changed by 1 units.

$β\_{4}$ means the difference of log odds of death within the groups with LDL between 130-159 mg/dL when their LDL changed by 1 units.

$β\_{5}$ means the difference of log odds of death within the groups with LDL between 160-189 mg/dL when their LDL changed by 1 units.

$β\_{6}$ means the difference of log odds of death within and the groups with LDL higher than 190 mg/dL.

The intercept means the odds of death for people with LDL = 0 mg/dL. There is no scientific meaning because of there is no LDL =0 in real world.

* 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 test if 6 variables are equal to zero to see the linearity. The six variables are generated by mkspline. P-value <0.0001, that means we can reject the null hypothesis. There is strong evidence to show that is nonlinear.
	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.
* By Using fitted hazard ratio, we can get relative hazard 3.898429 for the group with LDL 70 mg/dL, 1.551728 for the group with LDL 70-99 mg/dL, 1.530498 for the group with LDL 100-129 mg/dL, 1.145805 for group with LDL 130-159 mg/dL, and 1.234703 for the group with LDL higher than 190 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?
* In homework 1-3, we only cut LDL to two groups- higher than 160 mg/dL or not, and didn’t take censoring data into account. The advantages in homework 4 and 5 is we can consider the censoring in the real word by using cox regression model, and we can see if what hazard is different in different LDL level by following the Mayo Clinic guideline to divide LDL to many groups in homework 5.
	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.
* Unlike HW4, we divided LDL in to 6 groups rather than a continuous variable. From HW4, we can see the trend that lower hazard ratio is with higher LDL in the figure, and the figures showed a slightly U-shaped. That is, people with extreme LDL is have higher hazarded. The same, we get the hazard ratio in problem 2 and 3 to prove the lowest relative hazard is for the group with LDL 160 mg/dL.
	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 prefer to use cox regression and keep LDL as a continuous variable. The advantage to use cox regression is we can consider the censoring problem in the real work, and provide more meaningful evidence to answer the scientific questions. However, we should use transform of the LDL to be the predictor because the previous homework showed the LDL is a lightly U- shaped, especially for the extreme LDL.

**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.