**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2014

**Homework #5**

February 3, 2014

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 is. 4 distinct groups (White, Black, Asian, other) are modeled with 4 regression parameters.

**Method**: The odds of subjects getting diabetes were compared among subjects with different races (White, Black, Asian, other) using a logistic regression model. Dummy variables for each race are generated. Since the model is saturated, we don’t have to do the robust regression. Statistical inference was based on the likely ratio statistic.

**Result**: The odds of diabetes for the group of white is 0.109, which is reported as baseline variable (intercept). The odds of diabetes for the group of black is 0.209, which is 93% higher than the odds of white (odds ratio is 1.929). The odds of diabetes for Asians is 0.0682 (37% lower than the odds of white) and odds of diabetes for the other races is 0.2 (84.3% higher than odds of white). Overall, according the p-value computed under likely ratio test (0.1096), we conclude there’s no significant association between the prevalence of diabetes and races.

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

In total, we have four parameters. Here’s our model:

 is the log odds of diabetes for the baseline value. In this question, it’s the group of white.

 (odds 0.109 is reported in STATA)

 is the log odds ratio of diabetes for group of black to group of white.

(odds ratio 1.929 is reported in STATA)

 is the log odds ratio of diabetes for group of Asian to group of white.

(odds ratio 0.628 is reported in STATA)

 is the log odds ratio of diabetes for group of others to the group of white.

(odds ratio 1.843 is reported in STATA)

*In STATA, the coefficients of logistic regression model is exponentiated*

* 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.

If ignore the multiple comparisons, we can conclude the odds of white is significantly different to 0 at 95% significance level (p-value for intercept < 0.05). Also, we can conclude that the odds ratio of black to white is significantly different to 1 (p-value for the first coefficient is 0.026 < 0.05). This is saying the odds of diabetes for black is different to the odds for white. We cannot say the odds ratio of diabetes for Asians to white is significantly different to 1, same for odds ratio for the other races to white (p-values are 0.448 and 0.437 respectively which are both greater than 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)?

The model is just the reparameterization of the model in a. So the inference is the same (since we can compute the same odds for each group), although the value of intercept and coefficients are changed.

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

In total, we have four parameters. Here’s our model:

 is the log odds of diabetes for the baseline value. In this question, it’s the group of black.

(odds 0.209 is reported in STATA)

 is the log odds ratio of diabetes for group of white to group of black.

(odds ratio 0.519 is reported in STATA)

 is the log odds ratio of diabetes for group of Asian to group of black.

(odds ratio 0.326 is reported in STATA)

 is the log odds ratio of diabetes for group of others to the group of black.

(odds ratio 0.956 is reported in STATA)

*In STATA, the coefficients of logistic regression model is exponentiated*

* 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 ignore the multiple comparisons, we can conclude the odds of black is significantly different to 0 at 95% significance level (p-value for intercept < 0.05), and the odds would be reasonable if the true odds is in between 0.126 and 0.348 at 95% confidence level. Also, we can conclude that the odds ratio of white to black is significantly different to 1 (p-value for the first coefficient is 0.026 < 0.05). This is saying the odds of diabetes for white is different to the odds for black. We cannot say the odds ratio of diabetes for Asians to black is significantly different to 1, same for odds ratio for the other races to black (p-values are 0.085 and 0.956 respectively which are both greater than 0.05).

* 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)?

Some people might think it’s okay remove those variables that show insignificant result but that would be wrong. We tend to ignore the interactions between variables. Thus, removing one single variable would have the values of the other variables changed in the model, which will make the interpretation be different or even much severe modeling structural problem.

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.

**Descriptive statistics**: The study consisted of 735 subjects who were followed for death from any cause for a Kaplan-Meier estimated average of 5.33 years, during which time 133 deaths were observed. Serum LDL measurements at the time of study enrollment were not available on 10 subjects, two of whom were observed to die after 0.189 and 0.657 years of observation, with the remaining subjects still alive after 5.05 to 5.91 years of observation. 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).

 Kaplan-Meier based table:

|  |  |  |
| --- | --- | --- |
|  | LDL Level at Study Enrollment (mg/dl) |  |
|  | 1-69 | 70-99 | 100-129 | 130-159 | 160-189 | 190-250 | Total |
| 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 Percentile 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 | NA | NA | NA | 5.54 y |

\*NA indicates that the corresponding percentile is not estimable with the available data.

According the table, 5-year survival probability for the lowest LDL group is only 59.1% whereas for the 160-189 LDL group is 88%. By looking at 10 percentile of survival, 160-189 LDL group has 90% of subjects survived at least 4.30 years whereas for 100-129 LDL group, 90% of subjects survived at least 3.41 years. Also we can tell on average, 80% of the subjects survived at least 5.54 year.



**Methods**: Distributions of time to death from any cause was compared across groups defined by serum LDL at baseline using proportional hazards regression modeling serum LDL as dummy variables. Dummy variables are generated using the categories suggested by the Mayo Clinic: 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 (See descriptive statistics or homework 1). The association of all cause mortality was summarized by the hazards ratio 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.

**Inferential results**: Data was available on 725 subjects having mean serum LDL of 126 mg/dL (SD 33.6 mg/dL; range 11 – 247 mg/dL). During an average of 5.33 years of observation, 131 of those subjects were observed to die. From a proportional hazards regression analysis, The baseline value (hazard for the lowest LDL group) is not reported. But we can tell that the hazard of 70 - 99 LDL group is 60.2% lower (hazard ratio 0.398) comparing to the baseline group. For 100 – 129, 130 – 159, 160 – 189 and 190 – 250 groups, the hazards are 60.7%, 70.6%, 74.4% and 68.3% lower respectively compare to the baseline group. We found the wald based chi-square statistics is 15.42 and the 2-sided p-value is 0.0087. This suggests that we can reject the null hypothesis that the risk of death from any cause is not associated with serum LDL levels in favor of a tendency for lower mortality with higher serum LDL levels with 95% significance level.

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

In total, we have 6 parameters: intercept, β1, β2, β3, β4, β5.

For PH regression model, usually we don’t look at the baseline hazard, which is the intercept. However, when applying dummy variable model, the intercept might be meaningful because it’s the log hazard of the lowest LDL group.

β1 is the log hazard ratio for 70-99 LDL group to baseline.

 (hazard ratio 0.398 is reported in STATA)

β2 is the log hazard ratio for 100-129 LDL group to baseline.

(hazard ratio 0.393 is reported in STATA)

β3 is the log hazard ratio for 130-159 LDL group to baseline.

(hazard ratio 0.294 is reported in STATA)

β4 is the log hazard ratio for 160-189 LDL group to baseline.

(hazard ratio 0.257 is reported in STATA)

β5 is the log hazard ratio for 190-250 LDL group to baseline.

(hazard ratio 0.317 is reported in STATA)

*\*(When using cox regression command in STATA, it reports the exponentiated values)*

* 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 perform a F-test or Chi-square test to test if there’s non-linear relationship between variables (We used the dummy variables, but our main interest is always the association between instantaneous rate of death and .LDL level). Null Hypothesis: There’s no non-linear relationship between instantaneous rate of death and .LDL level. Alternative hypothesis: There is a non-linear association between instantaneous rate of death and LDL level. According to the result provided by STATA, p-value of the test is 0.3988 > 0.05. Thus, we failed to reject the null hypothesis which means we cannot say there exists a non-linear relationship between instantaneous rate of death and LDL level. So our model is good.

* 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.

see the graph in 4 b.

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 question 2 a for descriptive statistics.

**Methods**: Distributions of time to death from any cause was compared across groups defined by serum LDL at baseline using proportional hazards regression modeling serum LDL as linear spline variables. Spline variables are generated using the categories suggested by the Mayo Clinic: : 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 (See descriptive statistics or homework 1). The association of all cause mortality was summarized by the hazards ratio 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.

**Inferential results**: Data was available on 725 subjects having mean serum LDL of 126 mg/dL (SD 33.6 mg/dL; range 11 – 247 mg/dL). During an average of 5.33 years of observation, 131 of those subjects were observed to die**. From a proportional hazard regression model of the hazards on serum LDL using linear splines, we find a statistically significant association between instantaneous risk of death and serum LDL (two-sided P < 0.05).** We estimate that the instantaneous risk of death for 0-70 LDL group is a relative 19.86% lower (hazard ratio 0.8014) for each 10 mg/dL higher serum LDL level baseline. Based on a 95% confidence interval, this observed hazard ratio suggesting lower death rates for groups of patients with higher LDL levels would not be judged unusual if the true instantaneous risk of death were anywhere from 3.61% to 33.4% lower in a group having baseline serum LDL 10 mg/dL higher than that in another group (95% CI for hazard ratio 0.6663 to 0.9639).

**\***Inference for the other groups is very similar so ignored.

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

According to the spline model, we have 6 parameters in total.

0-69 mg/dl LDL group: When comparing two groups with different serum LDL level in this group, the instantaneous risk of death is estimated to be 19.86% lower (hazard ratio 0.8014) for each 10 mg/dl difference in LDL level, with the group having the higher level of LDL tending toward a lower instantaneous risk of death.

70-99 mg/dl LDL group: When comparing two groups with different serum LDL level in this group, the instantaneous risk of death is estimated to be 18.52% lower (hazard ratio .8148) for each 10 mg/dl difference in LDL level, with the group having the higher level of LDL tending toward a lower instantaneous risk of death.

100-129 mg/dl LDL group: When comparing two groups with different serum LDL level in this group, the instantaneous risk of death is estimated to be 2.27% lower (hazard ratio .9773) for each 10 mg/dl difference in LDL level, with the group having the higher level of LDL tending toward a lower instantaneous risk of death.

130-159 mg/dl LDL group: When comparing two groups with different serum LDL level in this group, the instantaneous risk of death is estimated to be 3.67% higher (hazard ratio 1.03669) for each 10 mg/dl difference in LDL level, with the group having the higher level of LDL tending toward a higher instantaneous risk of death.

160-189 mg/dl LDL group: When comparing two groups with different serum LDL level in this group, the instantaneous risk of death is estimated to be 25.6% lower (hazard ratio .744) for each 10 mg/dl difference in LDL level, with the group having the higher level of LDL tending toward a lower instantaneous risk of death.

190-250 mg/dl LDL group: When comparing two groups with different serum LDL level in this group, the instantaneous risk of death is estimated to be 32.8% higher (hazard ratio 1.328) for each 10 mg/dl difference in LDL level, with the group having the higher level of LDL tending toward a higher instantaneous risk of death.

* 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 perform a F-test or Chi-square test to test if there’s non-linear relationship between variables (In this case, we used spline variables). Alternative hypothesis: There is a non-linear association between instantaneous rate of death and LDL level. Accroding to the result provided by STATA, p-value of the test is 0.0788 > 0.05. Thus, we failed to reject the null hypothesis which means we cannot say there exists a non-linear relationship between instantaneous rate of death and LDL level. So our model is good.

* 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.

see the graph in 4 b

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?

Regression models we have been used are essentially based on proportional hazard regression (hw4 is about transformation and hw5 is about dummy and spline variables). By using proportional hazard regression, we have all the information of the censored data so less information is lost comparing to the dichotomized model used in hw 1-3.

* 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.



Similarity: Generally, there’s a downward trend for all the fitted lines in hw4 and 5 as LDL increases. Also, since they are all relative hazard, the hazard ratio is all 1 at LDL =160.

Difference: Fitted line for dummy variable is not continuous. It’s split into 5 segments and the hazard ratio only varies from segment to segment. Fitted line for spline variable looks like continuous. It captures subtler changing of hazard as LDL changed with respective to the other fitted lines, although the general trend for all of them is going downward.

* 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 choose spline model.

First of all, dummy variable model is not preferred because it’s hard to tell the association between two continuous variables through categorical model unless the data is clustered and the association is very obvious.

Second, compare to the transformed model, spline model definitely seemed to have more flexible and fitted. It has similar characters as the lowess curve. Since it’s able to capture the subtle association between variables, we could tell the association in each strata in a better way with respective to curves or line generated under the transformed models.

Also, if we were to predict the hazard ratio for any certain LDL level from 0 to 250 mg/dl, spline model might give us a more accurate prediction (borrow information) since it’s more sensitive in each strata.

**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.