**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2014

**Homework #5**February 10, 2014

1. **Methods:** A robust logistic model was fitted to determine the association between diabetes and race. Dummy variables for each race category, except the reference group, were created. A Wald test was preformed to test the overall null hypothesis that there is no association between diabetes and race. The odds ratio comparing odds of diabetes for each race group to white subjects was computed. A second model was run to compute the odds of the prevalence of diabetes for each race group compared to black subjects. For all odds ratios (the 95% CI is calculated for log odds and then transformed -1) computed, 95% confidence intervals were calculated using robust standard error estimates. In both models pairwise tests of the null hypothesis were preformed to determine if there is a statistically significant difference in odds of diabetes between each race group and the reference group (you cannot interpret the individual p-values and CIs due to multiple comparison -1). All tests were two-sided and evaluated at the 0.05 alpha level (This is not correct, the alpha level is not 0.05 due to multiple comparison -1). Missingness needs to be reported (-1).  
   1. A robust logistic model was fitted to determine the association between diabetes and race. Dummy variables for each race category, except the reference group (indicate which is the reference group), were created. A Wald test was preformed to test the overall null hypothesis that there is no association between diabetes and race. (see the comments above)  
        
      The model is saturated. The four groups being modeled are different race groups. The model contains four parameters, diabetes indicator (this one is the response variable, the intercept is the fourth parameter) and three dummy variables that indicate race. The number of groups and the number of parameters are the same. The p-value yielded from the Wald test is 0.0956 (the p-vlaue seems wrong 0.110 -1). Therefore, we cannot reject the null hypothesis that there is an association between diabetes and race. (you also need to report descriptive statistics such as # of people in each group and the observed prevalence of diabetes -2; please report the individual odds ratios for each group and interpret them carefully considering multiple comparison issue -1)

**Grade: 2/10.** See the comments interweaved in your answer. You really need to read the answer key carefully to learn how to make a comprehensive inferential results report. Be careful about multiple comparison.

* 1. The coefficients estimated are the odds ratios comparing the odds of diabetes between each race group and white subjects (not for the intercept -1). The odds of diabetes is 1.929 times higher for black subjects compared to white subjects. The odds of diabetes is 0.6282 times higher (as high as) for Asians compared to white subjects (This sentence seems wrong -1). The odds of diabetes is 1.843 times higher for other race groups compared to white subjects. The (exponentiated -1) intercept can be thought of odds of diabetes for the reference group, black subjects (I thought the reference group is the Whites? -1). The odds of diabetes are 0.1085 for white subjects.

Grade: 1/3

* 1. Only the p-value comparing the odds of diabetes between black and white subjects was less than 0.05 (p-value of 0.026). Therefore, we can reject the null hypothesis and say with some certainty that there is a difference in the odds of diabetes between black and white race groups. The p-values yielded from the test of the null hypothesis comparing Asians to white subjects was 0.449, and the p-value yielded from the test comparing other race groups to white subjects was 0.438. In both of these cases we fail to reject the null hypothesis and do not have enough evidence to say there is a difference in odds of diabetes between these groups. The p-value associated with testing the null hypothesis that the odds of diabetes for white subjects is 1 (meaning the risk of diabetes is 50% ) is less than 0.001. Therefore, we reject the null hypothesis and can say the odds of diabetes for white subjects is not one (it’s more intuitive to use log odds =0 -1).

Grade: 3/3

* 1. The inference found in (a) is the same found with this model. The p-value yielded from the Wald test assessing the association between diabetes and race is the same. The model fitted is a reparameterization of the one fitted in (a). The reference group is now black subjects, instead of white subjects so the odds ratios computed are comparing the odds of diabetes in each race group to black subjects. The overall test of association is the same, but the coefficients estimates will be different.

Grade: 3/3

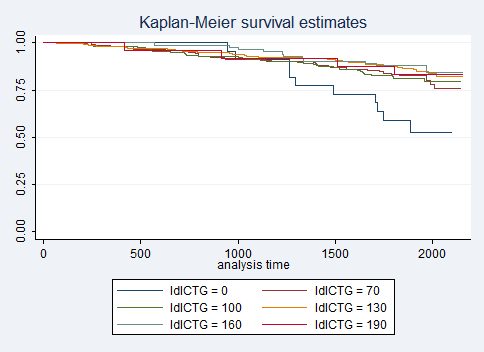
* 1. The coefficients estimated are the odds ratios comparing the odds of diabetes between each race group and black subjects (not for the intercept -1). The odds of diabetes are 49.15% lower for white subjects compared to black subjects. The odds of diabetes are 67.42% lower for Asians compared black subjects. The odds of diabetes are 4.444% lower for other race groups compared to black subjects. The intercept (the exponentiated intercept -1) can be thought of odds of diabetes for the reference group, black subjects. The odds of diabetes for blacks are 0.2093.

Grade: 1/3

* 1. Only the p-value comparing the odds of diabetes between white and black subjects was less than 0.05 (p-value of 0.026). Therefore, we can reject the null hypothesis and say with some certainty that there is a difference in the odds of diabetes between white and black race groups. The p-values yielded from the test of the null hypothesis comparing Asians to black subjects was 0.085, and the p-value yielded from the test comparing other race groups to black subjects was 0.956. In both of these cases we fail to reject the null hypothesis and do not have enough evidence to say there is a difference in odds of diabetes between these groups. The p-value associated with testing the null hypothesis that the odds of diabetes for black subjects is 1 (it’s more intuitive to use log odds =0) (meaning the risk of diabetes is 50%) is less than 0.001. Therefore, we reject the null hypothesis and can say the odds of diabetes for black subjects is not one.

Grade: 3/3

* 1. If we were to use the results found above in modeling building we would remove the dummy variables indicating Asian or other race from our model. This would be a poor decision. When looking at the pairwise tests we are actually preforming three statistical tests. For that reason, our Type I error is increased, and we are more likely to mistakenly reject the null hypothesis. This does seem the case, as the overall test conducted in (a) and (d) found that there was no relationship between race and diabetes, while we did find some significant relationships in parts (c) and (f). Secondly, although the same conclusions were reached regarding pairwise significance in (c) and (f) the p-values yielded were different (if you remove the indicator for Asian and others, you actually treating them in the same category as whites and thus the inference will become very different. This is major reason. This is not related to multiple testing really. Please refer to the key -3).

Grade: 2/5

* 1. **Methods**: LDL was recoded as a categorical variable based on the categories suggested by the Mayo Clinic. The levels are LDL less than 70 mg/dL, 70-99 mg/dL, 100-129 mg/dL, 130-159 mg/dL, 160-189 mg/dL, and LDL greater than 189 mg/dL. Dummy variables were coded to indicate which LDL category subjects belong to. Descriptive statistics were generated. Due to the censored nature of the data, Kaplan-Meier curves were plotted by each LDL category. Survival probabilities at 2 and 5 years for each LDL group were also calculated. A robust Cox proportional regression model was also fitted to assess the relationship between all-cause mortality and serum LDL levels, using the dummy variables mentioned above. Hazard ratios comparing instantaneous risk of death between LDL groups were estimated, as well as corresponding 95% confidence intervals (using robust standard error estimates). The reference group used in the model was a group with LDL serum levels less than 70 mg/dL (-1). An overall Wald test was performed to evaluate the association between LDL and all-cause mortality. Pairwise Wald tests were conducted to test the null hypothesis that risk of death is the same between subjects with LDL levels less than 70 mg/dL and other LDL groups (The pairwise tests are very bad.-1). All tests were two-sided and preformed at the 0.05 alpha level (you alpha level will be greater than 0.05 due to multiple comparison -1). (Also, your methods for descriptive statistics are insufficient -1 )

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Serum LDL Categories (mg/dL) | | | | | |
|  | < 70 | 70–99 | 100-129 | 130-159 | 160-189 | >189 |
| N Subjects | 22 | 143 | 228 | 225 | 83 | 24 |
| N Death | 10 | 28 | 44 | 34 | 11 | 4 |
| 2-year Survival Prob. | 1 | 0.9580 | 0.9386 | 0.9556 | 0.9880 | 0.9583 |
| 5-year Survival Prob. | 0.5909 | 0.8322 | 0.8114 | 0.8711 | 0.8795 | 0.8333 |

Ten patients with missing serum LDL levels were removed from the sample populations. The Kaplan-Meier plot, shown above, displays the survival rates for each LDL groups for the study period. There is a lot of cross-over in the survival curves for each LDL group. A total of 133 deaths occurred during the study period. The breakdown of deaths by serum LDL level is shown in the table above. The two and five year survival probabilities for each LDL group are also presented in the table above. The survival probabilities tend to be higher in the moderately high LDL group (160-189 mg/dL). Also survival seems to drop as LDL increases, but increases again for LDL greater or equal to 130 mg/dL. (restricted mean of survival time, mean LDL etc. -1 )  
  
The p-value yielded from the two-sided Wald test, testing the null hypothesis that there is no association between serum LDL and all cause mortality is 0.0087. We, therefore, reject the null hypothesis and can say with some certainty there is an association between serum LDL level and all cause mortality (This inference is far from sufficient. Please refer to the key. -2).

Grade: 2/10

* 1. The coefficients reported in the model are the hazard ratios (not for the intercept) comparing risk of death between each LDL group and the reference group, subjects with LDL less than 70 mg/dL. Subjects with serum LDL levels between 70 and 99 mg/dL have a 60.196% lower instantaneous risk of death compared to subjects with serum LDL levels less than 70 mg/dL. Subjects with serum LDL levels between 100 and 129 mg/dL have a 60.74% lower instantaneous risk of death compared to subjects with serum LDL levels less than 70 mg/dL. Subjects with serum LDL levels between 130 and 159 mg/dL have a 70.61% lower instantaneous risk of death compared to subjects with serum LDL levels less than 70 mg/dL. Subjects with serum LDL levels greater than 189 mg/dL have a 68.33% lower instantaneous risk of death compared to subjects with serum LDL levels less than 70 mg/dL. The intercept can be thought of as risk of death for the reference group being compared to itself, so it is always 1 (it is the baseline hazard function for the reference group -1).

Grade: 4/5

* 1. To assess whether the model fitted above provides a better fit than the model that only uses a continuous variable we would need to include all the dummy variables and continuous LDL into a model and simultaneously test if all the dummy variables fit lines with slopes equal to zero. That is, we would have to test of all the estimated parameters for the dummy variables are equal to 0, when included in a model with continuous LDL. Using a Chi-squared test to assess this, we find a p-value of 0.3988. We cannot reject the null hypothesis that the coefficients of the dummy variables are equal to 0. That is, we cannot say that a model that only includes the dummy variables (possibly a non-linear model) provides a better fit compared to the model uses LDL continuously.

Grade: 5/5

* 1. The hazard rates for each LDL category were computed relative to subjects with serum LDL levels of 160 mg/dL by computing hazard ratios for all subjects and dividing by the estimated hazard ratio for subjects with serum LDL of 160 mg/dL
  2. **Methods:** Due to the censored nature of the data, Kaplan-Meier curves were plotted by each LDL category. Survival probabilities at 2 and 5 years for each LDL group were also calculated. A robust Cox proportional hazards model was fitted to assess the relationship between all cause mortality and serum LDL levels, where serum LDL was fit as linear splines. Within each LDL category as defined by the Mayo Clinic a Cox proportional hazard model was fitted to estimate the hazard ratio comparing two subjects with a one unit difference in serum LDL level. An overall Wald test was conducted to test the association between LDL and all cause mortality. Wald tests were also conducted to determine if within LDL categories hazard rates were the same between subjects with a 1 mg/dL difference in serum LDL levels. Finally, a likelihood ratio test was preformed to determine if modeling LDL as splines resulted in a better fit compared to modeling LDL as a continuous variable (I don’t think part (a) ask you to do this).  
       
     The descriptive statistics describing the sample by each LDL group is shown above in (2).   
     p-value yielded from the Wald test of the overall null hypothesis that there is no association between LDL level and all cause mortality is less than 0.001. Therefore, we reject the null hypothesis and find that there is an association between LDL and all cause mortality. That is, for at least one of the LDL categories, the risk of death is different between subjects who have different LDL levels. Because there are six Wald tests being preformed to assess if there is an association between LDL and all cause mortality within each LDL group, a multiple testing issue arises. Instead, we are interested in testing if the hazard ratio estimates are the same within each LDL group. The results of this test are reported in (c).

Grade: 2/10. See comments for 2(a). The answer is far from sufficient.

* 1. The coefficients in this model represent the hazard ratio comparing risk of death between subjects with a 1 mg/dL difference in LDL levels, within each LDL group. Within the LDL less than 70 mg/dL group, there is a 2.190% decrease in risk of death for a 1 mg/dL increase in serum LDL. Within the LDL between 70 and 99 mg/dL group, there is a 2.027% decrease in risk of death for a 1 mg/dL increase in serum LDL. Within the LDL between 100 and 129 mg/dL group, there is a 0.2292% decrease in risk of death for a 1 mg/dL increase in serum LDL. Within the LDL between 130 and 159 mg/dL group, there is a 0.361% increase in risk of death for a 1 mg/dL increase in serum LDL. Within the LDL between 160 and 189 mg/dL group, there is a 2.908% decrease in risk of death for a 1 mg/dL increase in serum LDL. Within the LDL greater than   
     189 mg/dL group, there is a 2.8804% increase in risk of death for a 1 mg/dL increase in serum LDL. The intercept of this model can be thought of as the estimate for the instantaneous risk of death for the group where LDL is 0 (baseline hazard function for the reference group -1). However this is not scientifically meaningful, so it can be ignored.

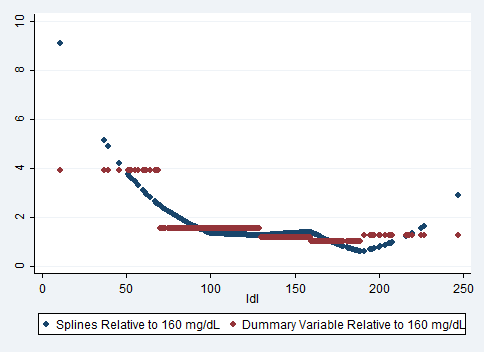
Grade: 4/5

* 1. To assess whether the model fitted above provides a better fit than the model that only uses a continuous LDL (i.e a linear model) we would need to test if the hazard ratio estimates for each LDL group are equal to each other using a Chi-squared test. The p-value yielded from this test is 0.0788. Therefore, we fail to reject the null hypothesis that the hazard ratio estimates are the same for each group. That is, we cannot say this model is better than a linear model where the independent variable is continuous.

Grade: 5/5

* 1. The hazard rates within each LDL cateogry were computed relative to subjects with serum LDL levels of 160 mg/dL by computing hazard ratios for all subjects and dividing by the estimated hazard ratio for subjects with serum LDL of 160 mg/dL
  2. In the methods in homeworks 4 and 5 we were not only able to test if there was an overall association between LDL and all cause mortality, but we were also able to test if there is a linear or non-linear association between LDL and all cause mortality. In homeworks 1-3 we had to work under the assumption the relationship between LDL and all cause mortality was linear (we also make this assumption in HW 4 when we model LDL continuously -1). Also, in earlier homeworks we often dichotomized death. In homeworks 4 and 5 we were able to use methods that accounted for the censored nature of the data and could treat observation time as a continuous variable. Finally, in homeworks 4 and 5 we were able to change the reference group when making comparisons in a multiplicative model (This is not sure, we can also change reference group in earlier HW by reparameterizing the LDL level). For example, we could treat subjects with LDL levels of 160 mg/dL as the reference group. We did not have this flexibility in earlier analyses (This is not ture).

Grade: 1/3

* 1. The results found in homework 4 and the one found in this model are fairly different. In the middle fitted in (2) the hazard ratio estimates comparing each LDL group were much larger than any estimates found in homework 4. From the model fitted in (3) the hazard ratio estimates are much lower than what was found in homework 4. The centered logged hazard ratio estimates found in homework 4 are similar to the hazard ratio estimates found in (3) with 160 mg/dL as the reference group. The dummy variable fitted hazard ratio estimates are similar to the centered hazard ratio estimates. The plots below show the similarities (I think you didn't answer the question on the point. Please refer to the key. You need to see all the lines in one graph instead of interpreting the difference in each graph separately, the estimates are actually pretty similar. The only difference is that splins and dummies are more flexible).  
       
     

Grade: 1/3

* 1. A priori, I would choose the analysis preformed in part (2) of this homework. As mentioned above, the advantages in this analysis are that we do not have to assume a non-linear association between LDL and all cause mortality (assuming linearity is not a bad practice. It is probably the best assumption to make when you know little. However, here due to multiplicative effect of predictors, log LDL may be the best). Although multiplicative models are sometimes more complicated to interpret, they may be more scientifically meaningful. This model may be especially scientifically meaningful as we considered pre-specified LDL groups that have scientific meaning. Although treating LDL as a continuous variable may increase statistical power, comparing groups that have a biological meaning may help us better answer the scientific question. (using flexible method for POI may not be a good choice and dummies are the last thing you would like to try)

Grade: 1/5 please refer to the key and see comments in your answer.