BIOSTAT 515/518 Winter 2014

Homework 5

ID 5256

1. (a) Methods: An association between the prevalence of diabetes and race was evaluated by comparing the odds of a diabetes diagnosis across groups defined by race modeled as a dummy indicator variables with whites as a reference group (for a total of 3 variables), using a logistic regression model with robust standard errors. Statistical inference on the two-sided p-values and were computed based on the Wald statistic.

Results: This is a saturated model, since we have equal number of predictors and groups. Of the 735 subjects with available data, 572 subjects were white (56 diabetic), 104 were black (18 diabetic), 47 were Asian (3 diabetic), and 12 were of other races (2 diabetic). The odds of being diagnosed with diabetes is for 0.109 for whites, 0.209 for blacks, 0.068 for Asians, and 0.200 for other races. The overall two-sided p-value of 0.0956 suggests we cannot with high confidence reject the null hypothesis that the odds of a diabetes diagnosis are not associated with race. It does not make sense to report the individual p-values and confidence intervals because they depend on the reference group specified and there are also other (4 choose 2) pairwise comparisons that need to be done to evaluate an association between race and diabetes.

(b) If we denote the model as (pi/(1-pi)) = exp(β0) x exp(β1Xi), then the intercept of the model is the odds of a diabetes diagnosis among whites. Thus the odds of a diabetes diagnosis in a group that is white is 0.109. The other coefficients are the odds ratios of a diabetes diagnosis in other racial groups compared to whites. So, the odds of a diabetes diagnosis is 1.93 times higher in blacks compared to whites, 0.628 times higher (or 0.312 times lower) in Asians compared to whites, and 1.84 times higher in other racial groups compared to whites.

(c) Using a 0.05 level of significance, the p-values reported in the regression output from part (a) would indicate a statistically significant difference of the odds ratio of a diabetes diagnosisbetween blacks and whites (p = 0.026), while a non-statistically significant difference of the odds ratio of a diabetes diagnosis between Asians and white (p = 0.449) and between other racial groups and whites (p = 0.438). There is statistical significance that the odds of a diabetes diagnosis for a group of whites is not 1 (p <0.0001). However there is not a real significance here since that is not our scientific question of interest.

(d) They would all be the same, since this is just a reparametrization of the model in (a). So:

Methods: An association between the prevalence of diabetes and race was evaluated by comparing the odds of a diabetes diagnosis across groups defined by race modeled as a indicator variables with blacks as a reference group (for a total of 3 variables), using a logistic regression model with robust standard errors. Statistical inference on the two-sided p-values and were computed based on the Wald statistic.

Inference: This is a saturated model, since we have equal number of predictors and groups. Of the 735 subjects with available data, 572 subjects were white (56 diabetic), 104 were black (18 diabetic), 47 were Asian (3 diabetic), and 12 were of other races (2 diabetic). The odds of being diagnosed with diabetes is for 0.109 for whites, 0.209 for blacks, 0.068 for Asians, and 0.200 for other races. The overall two-sided p-value of 0.0956 suggests we cannot with high confidence reject the null hypothesis that the odds of a diabetes diagnosis are not associated with race. It does not make sense to report the individual point estimates of odds ratios, individual p-values and confidence intervals because they depend on the reference group specified and there are also other (4 choose 2) pairwise comparisons that need to be done to evaluate an association between race and diabetes.

(e) The intercept of the model is the odds of a diabetes diagnosis among whites. Thus the odds of a diabetes diagnosis in a group that is black is 0.209. The other coefficients are the multiplicative difference in odds of diabetes in other racial groups compared to whites. So, the odds ratio of a diabetes diagnosisis 0.326 comparing Asians compared to blacks (Asians having lower odds of diabetes), 0.519 comparing whites to blacks (whites having lower odds of diabetes,, and 0.956 comparing other racial groups to blacks (other racial groups having lower odds of diabetes).

(f) Using a 0.05 level of significance, the p-values reported in the regression output from part (a) would indicate a statistically significant difference between the odds of a diabetes diagnosis between whites and blacks (p = 0.026), while a non-statistically significant difference between the odds of a diabetes diagnosis between Asians and blacks (p = 0.085) and between other racial groups and blacks (p = 0.956). There is statistical significance that the odds of a diabetes diagnosis for a group of blacks is not 1 (p <0.0001). However there is not a real significance here since that is not our scientific question of interest.

(g) If we use p values for individual regression parameters from a dummy variable regression to decide whether to include or exclude those variables in a regression model, we may decide that the difference in measured outcome in certain pairwise groups are statistically significant. For instance, in (c) and (f), we found that the odds ratio of a diabetes diagnosis between blacks and whites is statistically significant. However overall this is not the case. Having one pairwise statistical significance is not enough to determine that there is indeed a difference (in the odds of a diabetes diagnosis).

2. (a) Methods for descriptive statistics: Survival probabilities were categorized into 6 groups according to the Mayo Clinic guidelines: <70mg/dL, 70-99mg/dL, 100-129mg/dL, 130-159mg/dL, 160-189mg/dL, 190-250mg/dL. Within these categories, for all subjects with available data, Kaplan-Meier estimates of survival were calculated and graphed, and estimates of the 2 and 5 year survival probabilities, as well as the 10th and 20th percentiles of the survival distribution and the restricted mean survival during a period of observation that all LDL strata still had some subjects at risk (5.75 years). Number of subjects and number of deaths in each category was also reported.

Methods for inferential statistics: Distributions of time to death from any cause was compared across groups defined by serum LDL at baseline using robust cox proportional hazards regression (Breslow method for ties) modeling serum LDL as dummy variables using the categories suggested by the Mayo Clinic, using the group of <70mg/dL as the reference group. Quantification of association between all cause mortality was summarized by the two-sided p values computed using Wald chi-square 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.

Descriptive statistics results: Of the 735 subjects with available observations, 133 deaths were observed. The Kaplan-Meier estimated average survival time is 5.33 years (media 5.66 years, range 5.00 to 5.91 years). 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 observations. In the 725 subjects with available serum LDL measurements at enrollment, the mean LD was 126mg/dL (SD 33.6 mg/dL, range 11 to 247 mg/dL).

|  |  |  |
| --- | --- | --- |
|  | Serum LDL at Study Enrollment (mg/dL) | Total |
|  | < 70 | 70 – 99 | 100 – 129 | 130 - 159 | 160-189 | 190-247 |
| N subjects | 22 | 143 | 228 | 225 | 83 | 24 | 725 |
| N deaths | 10 | 28 | 44 | 34 | 11 | 4 | 131 |
| 2-year surv prob | 100% | 95.8% | 93.9% | 95.6% | 98.8% | 95.8% | 96.7% |
| 5-year surv prob | 59.1% | 83.2% | 81.1% | 87.1% | 88.0% | 83.3% | 86.0% |
| 10th percentile of survival | 3.46y | 3.80y | 3.41y | 4.30y | 4.53y | 4.13y | 3.66y |
| 20th percentile of survival | 3.55y | 5.44y | 5.36y | NA | NA | NA | 5.54y |
| 5.75 year restricted mean of survival | 4.91y | 5.24y | 5.23y | 5.35y | 5.45y | 5.32y | 5.29y |



Inferential results: Data was available on 725 subjects having mean serum LDL of 126mg/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, a two-sided p-value of 0.0087 suggests that we can with high confidence reject the null hypothesis that all-cause mortality is not associated with serum LDL levels. However, from this model, we cannot directly understand what kind of association it is.

(b) The intercept of the model is the baseline hazard of a group with serum LDL levels below 70mg/dL, some complicated function, and is actually not outputted in STATA unless specified. The other parameters are hazard ratios of the categorized groups compared to baseline. A group defined by serum LDL levels between 70mg/dL and 99mg/dL has a hazard ratio of 0.398 compared to a group of <70mg/dL. A group defined by serum LDL levels between 100mg/dL and 129mg/dL has a hazard ratio of 0.393 compared to baseline. A group defined by serum LDL levels between 130mg/dL and 159mg/dL has a hazard ratio of 0.294 compared to baseline. A group defined by serum LDL levels between 160mg/dL and 189mg/dL has a hazard ratio of 0.257 compared to baseline. A group defined by serum LDL levels above 190mg/dL has a hazard ratio of 0.317 compared to baseline.

(c) To assess whether the regression model used in this problem provides a “better fit” than a model that uses only a continuous linear term for LDL, a linear term is added to this model and then the dummy variables are tested together to see if they were zero.

Results of this analysis: The two-sided chi-square p-value is 0.3988. Thus from a test for nonlinearity based on the categorical dummy variables in the proportional hazard regression, we cannot with confidence reject the null hypothesis that true association between death from any cause and serum LDL is not adequately described by a hazard function that is linear in LDL. So including the dummy variables does not give us more information beyond a linear model.

(d) For each group, the hazard ratio relative to a group having serum LDL of 160mg/dL was computed. Results presented in Q4.

3. (a) Descriptive statistics methods & results: See question 2(a). They are exactly the same.

Methods for inferential statistics: Distributions of time to death from any cause was compared across groups defined by serum LDL at baseline using modeling serum LDL as dummy variables using the categories suggested by the Mayo Clinic. The groups are defined by linear splines: <70mg/dL, 70-99mg/dL, 100-129mg/dL, 130-159mg/dL, 160-189mg/dL, 190-250mg/dL. Quantification of association between all cause mortality was summarized by the hazards ratios of each spline computed from the regression models, with confidence intervals and two-sided p values computed using Wald chi-square 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 126mg/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, we estimate that the hazard ratios are 0.978 for each 1mg/dL difference in serum LDL levels for groups with serum LDL levels below 70mg/dL (those with higher serum LDL tending to lower hazards), 0.980 for each 1mg/dL difference in serum LDL levels for groups with serum LDL levels between 70mg/dL and 99mg/dL (those with higher serum LDL tending to lower hazards), 0.998 for each 1mg/dL difference in serum LDL levels for groups with serum LDL levels between 100mg/dL and 129mg/dL (those with higher serum LDL tending to lower hazards), 1.004 for each 1mg/dL difference in serum LDL levels for groups with serum LDL levels between 130mg/dL and 159mg/dL (those with higher serum LDL tending to higher hazards), 0.971 for each 1mg/dL difference in serum LDL levels for groups with serum LDL levels between 160mg/dL and 189mg/dL (those with higher serum LDL tending to lower hazards), and 1.03 for each 1mg/dL difference in serum LDL levels for groups with serum LDL levels above 190mg/dL (those with higher serum LDL tending to higher hazards). The individual p-values and confidence intervals are not readily interpretable. An overall two-sided p-value of <0.0001 suggest that we can with high confidence reject the null hypothesis that all-cause mortality is not associated with serum LDL levels. However, from this model, we cannot directly understand what kind of association it is.

(b) The intercept of the model is the baseline hazard, i.e. instantaneous risk of death for a group serum LDL is 0mg/dL, some complicated function, and is actually not outputted in STATA unless specified. Since this is not scientifically relevant, I am not reporting the value. The other parameters are hazard ratios of a difference of 1mg/dL in serum LDL of observations between two knobs. Thus the hazard ratios are 0.978 for each 1mg/dL difference in serum LDL levels for groups with serum LDL levels below 70mg/dL (those with higher serum LDL tending to lower hazards), 0.980 for each 1mg/dL difference in serum LDL levels for groups with serum LDL levels between 70mg/dL and 99mg/dL (those with higher serum LDL tending to lower hazards), 0.998 for each 1mg/dL difference in serum LDL levels for groups with serum LDL levels between 100mg/dL and 129mg/dL (those with higher serum LDL tending to lower hazards), 1.004 for each 1mg/dL difference in serum LDL levels for groups with serum LDL levels between 130mg/dL and 159mg/dL (those with higher serum LDL tending to higher hazards), 0.971 for each 1mg/dL difference in serum LDL levels for groups with serum LDL levels between 160mg/dL and 189mg/dL (those with higher serum LDL tending to lower hazards), and 1.03 for each 1mg/dL difference in serum LDL levels for groups with serum LDL levels above 190mg/dL (those with higher serum LDL tending to higher hazards).

(c) To assess whether the regression model used in this problem provides a “better fit” than a model that uses only a continuous linear term for LDL, since a straight line is a special case of linear splines, all the parameters (individual slopes between knobs) are tested together to see if the difference is 0.

Results of this analysis: The two-sided chi-square p-value is 0.0788. Thus from a test for nonlinearity based on the dummy variables from linear splines in the proportional hazard regression, we cannot with confidence reject the null hypothesis that true association between death from any cause and serum LDL is not adequately described by a hazard function that is linear in LDL. So including the linear splines variables does not give us more information beyond a linear model.

(d) For each group, the hazard ratio relative to a group having serum LDL of 160mg/dL was computed. Results presented in Q4.

4. (a) In HW1, both mortality and outcome were modeled as binary predictor variables in various ways. In HW2, the mean serum LDL of groups defined by vital status at 5 years were computed in various ways. In HW3, serum LDL was dichotomized into a binary variable and the probability and odds of death were computed. None of these took into account the survival times, whereas in HW4 and W5 we included time as a variable when evaluating the association between mortality and serum LDL. Thus we gain more precision. Furthermore the analyses in HW5 ignore the order of predictor of interest and so allow us to detect nonlinear trends. The linear splines model does not assume a step function, thus allow us to detect nonlinear trends with better precision than pure categorization; however there is some loss of precision due to categorization, compared to modeling serum LDL as a continuous variable.

(b)



From the above graph, all models seem to predict a general downward trend in hazard ratios. The fitted relative hazard ratios from HW4 all appear to look quite similar. The quadratic fit estimates the lowest hazard ratios for LDL levels in the middle and higher ratios at the two extreme ends of LDL levels, compared to other fits. The linear and log fits predict a general downward trend. However we cannot conclude that the true trend in hazard ratios are indeed log or quadratic from this graph, since the hazard ratios need to be log or quadratic across the entire real line. The linear splines fitted values are similar to the linear and log models, except it is not smooth (several “knobs” present, as expected). The categorical fit has a “choppy” step function and does not seem very useful to be used for intrapolation to unobserved groups (i.e. borrowing data).

(c) A priori, I would perform a cox proportional hazards regression on ldl defined as a log transformed continuous variable to explore an association between all cause mortality and serum LDL in an elderly population. Dividing a continuous variable into categories loses precision, and there really isn’t any scientific reason to do that. It makes more scientific sense to indicate the predictor as serum LDL and the outcome as mortality. Since we know this dataset is censored, it makes sense to use cox proportional hazards regression methods rather than other kinds of regressions. Lastly, since there is biological evidence to measure LDL levels on the log scale, a log transformed continuous LDL variable makes the most sense as it is interpretable.