**Biost 515/518**

**Homework #5**

1. Perform a statistical regression analysis evaluating an association between prevalence of diabetes and race by comparing the odds of a diabetes diagnosis across groups defined by race.
   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.

**Methods: A logistic regression model was used to evaluate an association between prevalence of diabetes and race by comparing the odds of a diabetes diagnosis across groups defined by race. Because race is a nominal variable, using whites as the reference group, indicator variables were created for black (black=1, all other races=0), Asian, and Other. This regression model is saturated because we are comparing four groups (white, black, Asian, and other subjects) using four parameters (the intercept and three slope parameters). The Huber-White sandwich estimator of the standard error was used. Statistical inference was based on the Wald statistic computed from the regression slope parameters and their standard errors, with two-sided p-values and 95% confidence intervals computed using the approximate normal distribution for logistic regression parameter estimates. The overall p-value was computed using a Wald chi-square statistic with 3 degrees of freedom.**

**Inference: Data was available regarding race and diabetes diagnoses for all 735 subjects in the dataset, of whom 572 were white, 104 were black, 47 were Asian, and 12 were of a different racial group (“Other”). The odds of a diabetes diagnosis among white subjects was 0.109, the odds among black subjects was 0.209, the odds among Asian subjects was 0.0682, and the odds among subjects of another racial group was 0.200. Since the model is saturated, the regression parameter estimates of the odds ratios exactly match the observed odds ratios. These odds ratios are 1.93 for blacks to whites, 0.628 for Asians to whites, and 1.84 for others to whites. An overall p-value of 0.0956 suggests that we do not have sufficient evidence to conclude that there is an association between race and the odds of a diagnosis of diabetes.**

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

**The intercept suggests that the estimated odds of a diabetes diagnosis among white subjects is 0.109. The slope parameter labeled “black” estimates the odds ratio between black subjects and white subjects as 1.929. The slope parameter labeled “Asian” estimates the odds ratio between Asian subjects and white subjects to be 0.628. Finally, the slope parameter labeled “other” estimates the odds ratio between people of other racial groups and whites to be 1.842.**

* 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.

**Ignoring multiple comparison concerns, based on a p-value of p < 0.001 for the intercept, we would conclude with high confidence that the odds of a diabetes diagnosis among white subjects is significantly different than 0. This is an unsurprising result. Further, based on a p-value of 0.026 for the slope parameter providing the odds ratio of black subjects to white subjects, we would reject the null hypothesis of equal odds among black and white subjects in favor of the alternative hypothesis that the odds of a diabetes diagnosis are greater among black subjects than white subjects. However, based on p-values of 0.449 for the odds ratios between Asians and whites and 0.438 for the odds ratio between other racial groups and whites, we would not have sufficient evidence to conclude that the odds of a diabetes diagnosis differ between Asian and white subjects or between subjects in other racial groups and white subjects.**

* 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 overall p-value in this case is (and must be) exactly the same as in part (a), since we are simply reparameterizing a saturated logistic regression model. The estimates and inference provided by this model can be exactly predicted from those reached with the first model in part (a), because they just represent linear transformations of the predictor variables and probabilities of complementary events sum to 1. Specifically, we can use the relationships defined below:**

**Let OW = odds among white subjects, OB = odds among black subjects, OA = odds among Asian subjects, and OO = odds among subjects of other racial groups. βij refers to parameter i in model j, where model 1 is the original model (whites as reference, then blacks, Asians, and others) and model 2 is the new model (blacks as reference, then whites, Asians, and others). Then β01 = OW, β11 = OB / OW, β21 = OA / OW, and β31 = OO / OW. Using these, we can express the coefficients of the new model as below.**

**β02 = OB = OW \* (OB / OW) = β01 \* β02**

**β12 = OW / OB = 1 / β11**

**β22 = OA / OB = (OA / OW) \* (OW / OB) = β21 / β11**

**β32 = OO / OB = (OO / OW) \* (OW / OB) = β31 / β11**

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

**The intercept suggests that the estimated odds of a diabetes diagnosis among black subjects is 0.209. The slope parameter labeled “white” estimates the odds ratio between white subjects and black subjects to be 0.519. The slope parameter labeled “Asian” estimates the odds ratio between Asian subjects and white subjects to be 0.628. Finally, the slope parameter labeled “other” estimates the odds ratio between people of other racial groups and whites to be 1.842.**

* 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.

**Ignoring multiple comparison concerns, based on a p-value of p < 0.001 for the intercept, we would conclude with high confidence that the odds of a diabetes diagnosis among black subjects is significantly different than 0. This is an unsurprising result. Further, based on a p-value of 0.026 for the slope parameter providing the odds ratio of white subjects to black subjects, we would reject the null hypothesis of equal odds among black and white subjects in favor of the alternative hypothesis that the odds of a diabetes diagnosis are lower among white subjects than black subjects (odds ratio estimate 0.519, which with 95% confidence is consistent with a true odds ratio between 0.291 and 0.925). However, based on p-values of 0.085 for the odds ratios between Asians and blacks and 0.956 for the odds ratio between other racial groups and blacks, we would not have sufficient evidence to conclude that the odds of a diabetes diagnosis differ between Asian and black subjects or between subjects in other racial groups and black subjects.**

* 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)?

**As demonstrated in (c) and (f) above, individual regression parameter p-values from a dummy variable regression change based on the parameterization of the model. Both models give exactly the same estimates, and in both the p-value for the odds ratio of the intercept and of blacks to whites (or vice versa) were significant (and the same) while the others were not statistically significant. However, the p-values for the second and third slope parameters changed dramatically between reparameterizations of the model; for example, the p-value for the odds ratio comparing Asians to whites was 0.449, while the p-value for the odds ratio comparing Asians to blacks was 0.085. Any method for deciding whether to include or exclude variables that bears the possibility of making a different decision based only on a reparameterization of the model, like this one does, is a dangerous method for such decision-making.**

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.
   1. Include full description of your methods, appropriate descriptive statistics, and full report of your inferential statistics.

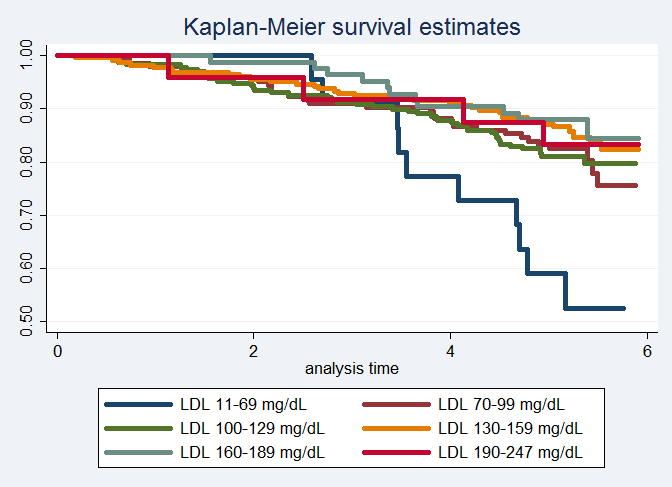
**Methods for Descriptive Statistics: Descriptive statistics for the censoring distribution included the minimum and maximum observed censoring times and the Kaplan-Meier estimates of the 10th, 50th, and 90th percentiles, as well as the mean time of follow-up, calculated as the area under the Kaplan-Meier estimate of the censoring distribution’s survivor curve.**

**Descriptive statistics for serum LDL levels included the number of cases with missing data. For cases with available data, the minimum, maximum, mean, standard deviation, and 25th, 50th, and 75th percentiles were presented. Serum LDL was stratified according to the Mayo Clinic guidelines, i.e., <70 mg/dL, 70-99 mg/dL, 100-129 mg/dL, 130-159 mg/dL, 160-189 mg/dL, and ≥ 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, the 10th and 20th percentiles of the survival distribution, and the restricted mean survival until 5.75 years (as all LDL categories still had subjects at risk until this point).**

**Descriptive Statistics: The study included 735 subjects who were followed for death from any cause for a Kaplan-Meier estimated average of 5.33 years (median 5.66 years, range 5.00 to 5.91 years), during which time 133 deaths were observed. Serum LDL measurements were not available on 10 subjects. Two of these subjects were observed to die after 0.189 and 0.657 years; the remainder were still alive after 5.05 to 5.91 observation years. Among the 725 subjects with data available, the mean serum LDL was 126 mg/dL, with standard deviation 33.6 mg/dL and range 11-247 mg/dL.**

**Estimates of the survival distribution within groups defined by serum LDL and in the entire sample (missing data omitted) are provided in the table below. Note that the NA means that the survival probability could not be calculated, and survival probabilities and percentiles were based on Kaplan-Meier estimates.**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Serum LDL at Baseline (mg/dL)** | | | | | | |
|  | **11 – 69** | **70 – 99** | **100 – 129** | **130 – 159** | **160 – 189** | **190 – 247** | **All** |
| **n Subjects** | 22 | 143 | 228 | 225 | 83 | 24 | 725 |
| **n Deaths** | 10 | 28 | 44 | 34 | 11 | 4 | 131 |
| **2 year Surv. Probability** | 100% | 95.8% | 93.9% | 95.6% | 98.8% | 95.8% | 96.7% |
| **5 year Surv. Probability** | 59.1% | 83.2% | 81.1% | 87.1% | 88.0% | 83.3% | 86.0% |
| **10th %ile Surv. Probability** | 3.46 yrs | 3.80 yrs | 3.41 yrs | 4.30 yrs | 4.53 yrs | 4.13 yrs | 3.66 yrs |
| **20th %ile Surv. Probability** | 3.55 yrs | 5.44 yrs | 5.36 yrs | NA | NA | NA | 5.54 yrs |

****

**Based on the plot above, we see the most significant divergence in the lowest LDL group (11-69 mg/dL), with rapidly decreasing survival estimates after 3 years. This is consistent with the estimates in the table, where survival among this lowest LDL group decreases to 59.1% by 5 years, compared to 80-90% in all of the other groups. We also note that overall two year and five year survival probabilities are highest among subjects with LDL 160-189 mg/dL, with estimates of 98.8% and 88.0%, respectively.**

**Methods: Serum LDL was divided into catgories as above, according to the Mayo Clinic guidelines, that is, <70 mg/dL, 70-99 mg/dL, 100-129 mg/dL, 130-159 mg/dL, 160-189 mg/dL, and ≥ 190 mg/dL. Indicator variables were created for each category, and the reference group was the group with serum LDL less than 70 mg/dL. Distributions of time to death from any cause was compared across these groups using Cox proportional hazards regression. An overall two-sided p-value was computed using Wald statistics based on the Huber-White sandwich estimator. Subjects missing data for serum LDL at baseline were omitted.**

**Inference: Data is 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 proportional hazards regression analysis using indicator variables for membership in each serum LDL category, we obtain an overall p-value of 0.0087. This p-value suggests that we can with high confidence reject our null hypothesis of no association between serum LDL category and instantaneous risk of death, concluding that at least one of the categories differs in instantaneous risk of death.**

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

**The intercept in the proportional hazards regression model from part (a) would be the baseline hazard among the group with serum LDL less than 70 mg/dL, but this is generally not estimated. The slope parameter labeled “70” estimates that the instantaneous risk of death among a group of subjects with LDL of 70-100 is 60.2% lower than the group with LDL less than 70 mg/dL (hazard ratio 0.398). The slope parameter labeled “100” estimates the hazard ratio between the group of subjects with LDL of 100-129 and the group with LDL <70, suggesting that the instantaneous risk of death is 60.7% lower among the group with higher LDL (hazard ratio 0.393). The slope parameter labeled “130” estimates the hazard ratio between the group of subjects with LDL of 130-159 and the group with LDL <70 to be 0.294, that is, that the instantaneous risk of death is 70.6% lower among the group with higher LDL. The slope parameter labeled “160” estimates the hazard ratio between the group of subjects with LDL of 160-189 and the group with LDL <70 to be 0.257, that is, that the instantaneous risk of death is 74.3% lower among the group with higher LDL. Finally, the slope parameter labeled “190” estimates the hazard ratio between the group of subjects with LDL of ≥190 and the group with LDL <70 to be 0.317, that is, that the instantaneous risk of death is 68.3% lower among the group with higher 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?

**I would test for linearity by fitting a Cox proportional hazards regression model using the indicator variables from the above analysis plus LDL treated as a continuous variable, then testing that all of the regression slopes for the indicator variables are 0. If all of these slopes are zero, that would suggest that the linear fit is appropriate. The p-value from this analysis is 0.399, suggesting that we do not have sufficient evidence to conclude that the true association between death from any cause and serum LDL is not adequately described by a log hazard function that is linear in LDL.**

* 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.

**No answer necessary; a variable was generated.**

1. Perform a statistical regression analysis evaluating an association between all-cause mortality and serum LDL 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, please see question 2 above.**

**Methods: Linear splines were fit within the ranges 0-69, 70-99, 100-129, 130-159, 160-189, and greater than or equal to 190 mg/dL serum LDL. Distributions of time to death from any cause was compared across serum LDL levels thus divided using proportional hazards regression. An overall two-sided p-value was computed using a chi-square test with 5 degrees of freedom using the Wald statistics based on the Huber-White sandwich estimator. Subjects missing data for serum LDL at baseline were omitted.**

**Inference: Data is 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 proportional hazards regression analysis using linear spline fits within each serum LDL category, we obtain an overall p-value < 0.0001. This p-value suggests that we can with high confidence reject our null hypothesis of no association between serum LDL category and instantaneous risk of death and conclude that instantaneous risk of death differs across serum LDL levels.**

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

**The intercept is the instantaneous risk of death for subjects with serum LDL of zero. This is not scientifically relevant, since 0 is not within the range of biologically plausible values for serum LDL. The slope parameter for ldl0, 0.978, is the estimated hazard ratio between two groups both with serum LDL between 0 and 69 mg/dL, but differing by 1 unit in serum LDL. The slope parameter for ldl70, 0.980, is the estimated hazard ratio between two groups both with serum LDL between 70 and 99 mg/dL, but differing by 1 unit in serum LDL. The slope parameter for ldl100, 0.998, is the estimated hazard ratio between two groups both with serum LDL between 100 and 129 mg/dL, but differing by 1 unit in serum LDL. The slope parameter for ldl130, 1.00, is the estimated hazard ratio between two groups both with serum LDL between 130 and 159 mg/dL, but differing by 1 unit in serum LDL. Note that we estimate that the instantaneous probability of death is nearly constant over this interval. The slope parameter for ldl160, 0.971, is the estimated hazard ratio between two groups both with serum LDL between 160 and 189 mg/dL, but differing by 1 unit in serum LDL. Finally, the slope parameter for ldl190, 1.03, is the estimated hazard ratio between two groups both with serum LDL greater than or equal to 190 mg/dL, but differing by 1 unit in serum LDL. Note that all of these parameters except the last are less than or equal to 1, indicating an overall negative trend in any relationship between serum LDL and instantaneous risk of death (that is, the groups with higher serum LDL within each category have lower 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?

**To test for nonlinearity, we perform a Wald test for the proportional hazard regression coefficients using the Huber-White sandwich estimator for the standard error. We test whether all of the spline slopes are equal; this would indicate that the splines are essentially fitting a single line. This test results in a p-value of 0.0788, indicating that at a 5% confidence level, we do not have sufficient evidence to conclude that the true association between death from any cause and serum LDL is not adequately described by a log hazard function that is linear in LDL.**

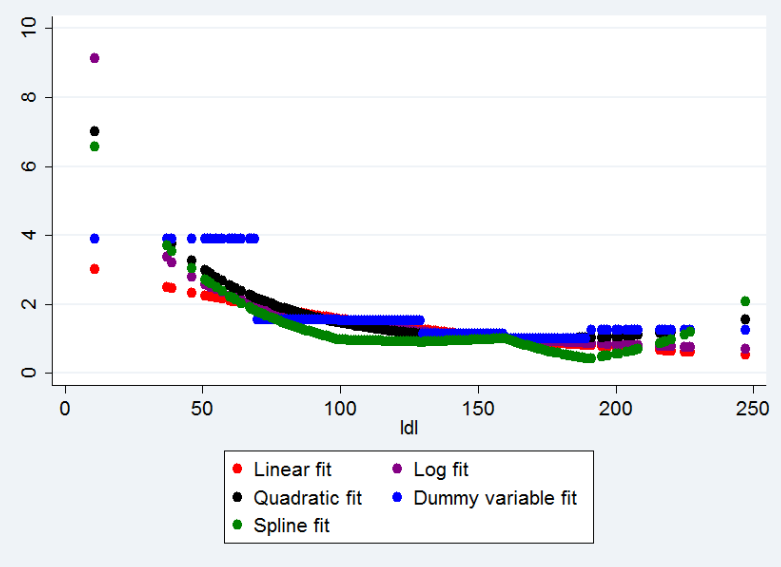
* 1. For each population defined by serum LDL value, compute the hazard ratio relative to a group having serum LDL of 160 mg/dL. (This will be used in problem 4). This can be effected by generating fitted hazard ratio estimates for each individual in the sample, and then dividing that fitted value by the fitted value for a subject having a LDL of 160 mg/dL.

**No answer necessary; a variable was generated.**

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 homeworks 4 and 5, time to death did not need to be dichotomized, since proportional hazards regression was used. The stratification of LDL required in the homework 5 analyses allowed more precision than the analyses in homeworks 1-3 that required the dichotomization of either LDL or death, though precision (“degrees of freedom”) was still lost compared to the models in homework 4, which treated LDL as a continuous variable. Homework 5 and the quadratic analysis in homework 4 allowed us to test for nonlinearity, while the identity and log-link transformations did not allow such tests.**

* 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.



**The figure above shows the fitted values for the five models fit in homeworks 4 and 5. Each model predicts a generally downward trend, although the spline fit is less monotonically decreasing than the other four. There are not significant differences between the linear, logarithmic, quadratic, and dummy variable fits in the middle range of the data; for low LDL values, the difference increases. The difference is especially apparent in the dummy variable fit, which, because it is a step function, predicts a high relative hazard over the entire range of 0 to 70 mg/dL LDL. However, over the rest of the range of serum LDL, the dummy variable fit follows the quadratic fit quite closely (though clearly not continuously, since it is a step function). The spline fit matches the linear and log fits from homework 4 well for low serum LDL levels, then drops below the other lines at two of its knots (100 and 190 mg/dL). Based on the results in the questions above, we cannot conclude that a linear fit does not adequately describe the relationship between log hazard and serum LDL, but there is remarkable similarity in the shape of the other four fits in the plot above. Of these four (log, quadratic, dummy variable, and spline), the logarithmic fit is most attractive, both because it is relatively easy to interpret and because it does not have jump points or knots, which limit generalizability of results.**

* 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?

***A priori*, I would have preferred to perform Cox proportional hazards regression using log transformed LDL treated as a continuous variable as the predictor, as in Homework 4 question (2). We have biological reason to believe the effect of LDL is on a multiplicative rather than additive scale, hence the log transformation. Testing for nonlinear associations is really first testing for an association and then testing for linearity, so it makes more sense to first just test for an association that is linear on the log scale. Then (in my previously specified analysis plan), if an association was found, I would use splines to test whether there is evidence to conclude that a linear fit does not adequately describe the association.**