**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.*

METHODS:

The odds of having diabetes was compared by racial status, defined as white vs. nonwhite subjects, using a logistic regression model. Statistical inference was based on the Wald statistic computed from the regression slope parameter and its standard error, with two-sided p value and 95% confidence interval computed suing the approximate normal distribution for logistic regression parameter estimates. As this model is saturated with two distinct groups (whites and nonwhites) and two regression parameters (slope and intercept), we did not use robust standard errors.

INFERENCE:
The study consisted of 735 subjects aged 65 and older. Out of 735 subjects included in this study, 572 or 77.8% were white and 164 (22.2%) subjects were “nonwhite.” The nonwhite group included black, Asian, and Hispanic subjects. Subjects were also defined by their diabetes status, with 656 subjects, or 89.3% having diabetes and 79 or 10.7% without a diagnosis of diabetes. Among the 572 (77.8%) white subjects, the odds of having diabetes is 33.94% less than it is for nonwhite subjects. The estimated odds ratio is 0.6606 with a 95% confidence interval of 0.3927 to 1.1113 and a p value that is not statistically significant of 0.118. It would not be unusual for the odds of diabetes among white patients to be between 39.27% to 111.13% of that of nonwhite patients. That is, to have an odds that is between 61.73% lower to 11.13% higher than their nonwhite counterparts of having a diagnosis of diabetes. Based on the p-value of 0.118, we fail to reject the null hypothesis that there is no association between race defined as white vs. nonwhite and 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).*

Odds ratio of diabetes = 0.6606 (race = white) + 0.1643

The slope of the logistic regression represents the difference in odds ratio between the white and the nonwhite group when there is a difference in one unit. Because this is a dummy variable, a one unit difference represents the two race groups, white vs. nonwhite. The intercept represents the odds ratio when x=0, or the odds ratio for the nonwhite group.

* 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.*

The results are not statistically significant as the p-value is 0.118. We fail to reject the null hypothesis that there is an association between race defined as white vs. nonwhite and a diagnosis of diabetes.

* 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 odds of having diabetes for black subjects is 95.58% higher than it is for nonblack subjects. The 95% confidence interval estimates that it would not be unusual for the odds of diabetes to be 10.32% higher to 246.71% higher for black subjects than it is for nonblack subjects, where being black is associated with higher risk for having diabetes. The confidence interval is 0.022 which is statistically significant with an α of 0.05%, suggesting that we can reject the null hypothesis that there is no relationship between being black and having a diagnosis of diabetes.

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

Odds ratio of diabetes = 1.9558 \*(race = black) + 0.1070

The slope of the logistic regression represents the difference in odds ratio between the black and the nonblack group when there is a difference in one unit. Because this is a dummy variable, a one unit difference represents the two race groups, black vs. nonblack. The intercept represents the odds ratio when x=0, or the odds ratio for the nonblack group.

* 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.*

With a p-value of 0.022 and a 0.05 level of significance, we can reject the null hypothesis that there is no association between being black and having a diagnosis of diabetes.

* 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)?*

Modeling with dummy variables ignores the order of the predictor of interest, in this case the racial group. That is, by grouping the nonwhite and nonblack categories into one group, we are ignoring differences that may exist within each group and lose power. Because of such different inferences that were drawn when only using one dummy variable, we can improve our prediction accuracy by including as much information in the model as possible. We could create dummy variables for each race so that the “zero” category was as specific as possible.

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.*

METHODS:

Distributions of time to death from any cause was compared across groups defined by serum LDL groups at baseline with the lowest LDL group (<70 mg/dL) as the referent group using proportional hazards regression modeling serum LDL as dummy variables of groups categorized by LDL levels suggested by the Mayo Clinic and reported in Homework #1. Quantification of association between 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 were omitted from analysis.

INFERENCE:

Data was available on 725 subjects having a mean serum LDL of 126 mg/dL (SD 33.6 mg/dL; range 11-247 mg/dL). There were 22 subjects (3.0%) in the group with LDL <70mg/dL, 143 subjects (19.7%) with LDL between 70-100 mg/dL, 228 subjects (31.5%) with LDL 100-130mg/dL, 225 subjects (31.0%) with LDL 130-160 (mg/dL), 83 subjects (11.5%) with LDL 160-190 mg/dL), and 24 subjects (3.3%) with LDL > 190 mg/dL. Over an average of 5.33 years of observation, 131 subjects were observed to die. From a proportional hazards regression analysis, we estimate that the instantaneous risk of death is 60.20% lower (hazard ratio 0.3980) among individuals who belong to the second lowest LDL group (70-100mg/dL) when compared to the lowest LDL group (<70 mg/dL). Based on a 95% confidence interval, this observed hazard ratio suggesting higher death rates for groups of patients with lower LDL levels would not be unusual if the true instantaneous risk of death was between 21.80% to 79.74% lower in the group with the second lowest LDL group (70-100 mg/dL). A two-sided p-value of 0.008 suggests that we can reject the null hypothesis that the risk of death from any cause is not associated with serum LDL levels when grouped by Mayo Clinic classification in favor of a tendency for higher mortality with lower serum LDL levels.

*LDL 100-130mg/dl*

We can estimate the risk for each Mayo Clinic LDL group. The estimated instantaneous risk of death for subjects with a serum LDL from 100-130 mg/dl is 60.74% lower (hazard ratio 0.3926) compared to the referent group. A 95% confidence interval estimates that this observed hazard ratio would not be unusual if the true instantaneous risk of death was between 25.58% to 79.29% less for participants with a serum LDL between 100-130mg/dL than it is for participants in the lowest LDL group (<70 mg/dl). A two-sided p-value of 0.004 suggests that we can reject the null hypothesis that the risk of death from any cause is not associated with serum LDL levels when grouped by Mayo Clinic classification in favor of a tendency for higher mortality with lower serum LDL levels.

*LDL 130-160mg/dl*

We can estimate the risk for each Mayo Clinic LDL group. The estimated instantaneous risk of death for subjects with a serum LDL from 130-160 mg/dl is 70.61% lower (hazard ratio 0.2939) compared to the referent group. A 95% confidence interval estimates that this observed hazard ratio would not be unusual if the true instantaneous risk of death was between 43.22% to 84.79% less for participants with a serum LDL between 130-160mg/dL than it is for participants in the lowest LDL group (<70 mg/dl). A two-sided p-value of <0.0001 suggests that we can reject the null hypothesis that the risk of death from any cause is not associated with serum LDL levels when grouped by Mayo Clinic classification in favor of a tendency for higher mortality with lower serum LDL levels.

*LDL 160-190mg/dl*

We can estimate the risk for each Mayo Clinic LDL group. The estimated instantaneous risk of death for subjects with a serum LDL from 160-190 mg/dl is 74.35% lower (hazard ratio 0.2565) compared to the referent group. A 95% confidence interval estimates that this observed hazard ratio would not be unusual if the true instantaneous risk of death was between 42.01% to 88.65% less for participants with a serum LDL between 160-190mg/dL than it is for participants in the lowest LDL group (<70 mg/dl). A two-sided p-value of 0.001 suggests that we can reject the null hypothesis that the risk of death from any cause is not associated with serum LDL levels when grouped by Mayo Clinic classification in favor of a tendency for higher mortality with lower serum LDL levels.

*LDL >190mg/dl*

We can estimate the risk for each Mayo Clinic LDL group. The estimated instantaneous risk of death for subjects with a serum LDL >190 mg/dl is 68.33% lower (hazard ratio 0.3167) compared to the referent group. A 95% confidence interval estimates that this observed hazard ratio would not be unusual if the true instantaneous risk of death was between 1.08% to 89.86% less for participants with a serum LDL between >190 mg/dL than it is for participants in the lowest LDL group (<70 mg/dl). A two-sided p-value of 0.048 suggests that we can reject the null hypothesis that the risk of death from any cause is not associated with serum LDL levels when grouped by Mayo Clinic classification in favor of a tendency for higher mortality with lower serum LDL levels. However, the overall inverse trend between serum LDL and mortality risk appears to level off or reverse when values of LDL are greater than 190mg/dL, as our hazard ratio point estimate is closer to 1 for this LDL group than it was for the groups with LDLs between 130-190 mg/dL.

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

Instantaneous risk of death compared to LDL 1 group = 0.3980 (LDL 2 group) + 0.3926 (LDL group 3) + 0.2939 (LDL 4 group) + 0.2565 (LDL 5 group) + 0.3167 (LDL 6 group)

The slopes correspond to the hazard ratio for each LDL group when it equals one relative to the referent group, the LDL 1group (LDL< 70mg/dL). Because these are dummy variables, the other terms will drop out as they have values of zero.

* 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 will perform a scatter plot of each hazard regression model with Lowess fitted line to compare the fit of each model. These methods are purely qualitative, but we can see a u-shaped relationship in model in Figure 2 where LDL is a categorical variable. This is not observed in Figure 3 where the relationship looks curvilinear. Also, although it is an imperfect measure, the p-values for the chi-squared test for each model is 0.0087 for the model with LDL as a categorical variable, and slightly higher at 0.0093 when LDL is modeled continuously, suggesting that the first model may be a slightly better fit.



Figure 1



Figure 2



Figure 3

* 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.*

This was performed.

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.*

METHODS:

Distributions of time to death from any cause was compared across groups defined by serum LDL groups at baseline using proportional hazards regression modeling serum LDL using linear splines split into groups based on serum LDL levels as defined by the Mayo Clinic (and discussed in Homework 1). Quantification of association between all cause mortality was summarized by the hazards ratio computed from the regression model, with confidence intervals and two-sided p values of 0.05 computed using Wald statistics based on the Huber-White sandwich estimator. Subjects missing data for serum LDL were omitted from analysis.

INFERENCE:

Data was available on 725 subjects having a mean serum LDL of 126 mg/dL (SD 33.6 mg/dL; range 11-247 mg/dL). There were 22 subjects (3.0%) in the group with LDL <70mg/dL, 143 subjects (19.7%) with LDL between 70-100 mg/dL, 228 subjects (31.5%) with LDL 100-130mg/dL, 225 subjects (31.0%) with LDL 130-160 (mg/dL), 83 subjects (11.5%) with LDL 160-190 mg/dL), and 24 subjects (3.3%) with LDL > 190 mg/dL. Over an average of 5.33 years of observation, 131 subjects were observed to die.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| LDL group | N (total=725) | Hazard Ratio, per 10mg/dl unit increase | p-value | 95% confidence interval |
| <70mg/dL | 22 | 0.8014 | 0.019 | 0.6662 - 0.9636 |
| 70 to <100mg/dL | 143 | 0.8146 | 0.139 | 0.6231 - 1.0691 |
| 100 to <130mg/dL | 228 | 0.9772 | 0.835 | 0.7875 – 1.2130 |
| 130 to <160mg/dL | 225 | 1.0366 | 0.773 | 0.8121 – 1.3245 |
| 160 to <190mg/dL | 83 | 0.7443 | 0.181 | 0.4829 – 1.1469 |
| >190mg/dL | 24 | 1.3283 | 0.261 | 0.8096 – 2.1790 |

*LDL <70mg/dL*

 From a proportional hazards regression analysis, we estimate that the instantaneous risk of death is 2.19% lower (hazard ratio 0.9781) for every 1 mg/dL increase in serum LDL among individuals who belong to the lowest LDL group (<70 mg/dL). For every 10 mg/dL increase in serum LDL, there is a decrease in instantaneous risk of death of 19.86% (hazard ratio 0.8014) for subjects with a baseline serum LDL < 70mg/dL. Based on a 95% confidence interval, this observed hazard ratio suggesting higher death rates as LDL increases would not be unusual if the true instantaneous risk of death was between 3.64% to 33.38% lower for every 10 mg/dL increase in baseline serum LDL. A two-sided p-value of 0.019 suggests that we can reject the null hypothesis that the risk of death from any cause is not associated with serum LDL levels when grouped by Mayo Clinic classification in favor of a tendency for higher mortality with lower serum LDL levels.

*LDL 70 – 100 mg/dL*

 From a proportional hazards regression analysis, we estimate that the instantaneous risk of death is 2.03% lower (hazard ratio 0.9797) for every 1 mg/dL increase in serum LDL among individuals who belong to the group with LDL between 70-100 mg/dL). For every 10 mg/dL increase in serum LDL, there is a decrease in instantaneous risk of death of 18.54% (hazard ratio 0.8146) for subjects with a baseline serum LDL 70-100 mg/dL. Based on a 95% confidence interval, this observed hazard ratio would not be unusual if the true instantaneous risk of death was between 37.69% lower to 6.91% higher for every 10 mg/dL increase in baseline serum LDL. A two-sided p-value of 0.139 suggests that we cannot reject the null hypothesis that the risk of death from any cause is not associated with serum LDL levels for subjects with a baseline LDL of 70-100mg/dL.

*LDL 100 – 130 mg/dL*

 From a proportional hazards regression analysis, we estimate that the instantaneous risk of death is 0.89% lower (hazard ratio 0.9977) for every 1 mg/dL increase in serum LDL among individuals who belong to the group with LDL between 100-130 mg/dL). For every 10 mg/dL increase in serum LDL, there is a decrease in instantaneous risk of death of 2.28% (hazard ratio 0.9772) for subjects with a baseline serum LDL 100-130 mg/dL. Based on a 95% confidence interval, this observed hazard ratio would not be unusual if the true instantaneous risk of death was between 21.25% lower to 21.30% higher for every 10 mg/dL increase in baseline serum LDL. A two-sided p-value of 0.835 suggests that we cannot reject the null hypothesis that the risk of death from any cause is not associated with serum LDL levels for subjects with a baseline LDL of 100-130 mg/dL.

*LDL 130 – 160 mg/dL*

 From a proportional hazards regression analysis, we estimate that the instantaneous risk of death is 0.36% higher (hazard ratio 1.0036) for every 1 mg/dL increase in serum LDL among individuals who belong to the group with LDL between 130-160 mg/dL). For every 10 mg/dL increase in serum LDL, there is an increase in instantaneous risk of death of 3.66% (hazard ratio 1.0366) for subjects with a baseline serum LDL 130-160 mg/dL. Based on a 95% confidence interval, this observed hazard ratio would not be unusual if the true instantaneous risk of death was between 19.79% lower to 32.45% higher for every 10 mg/dL increase in baseline serum LDL. A two-sided p-value of 0.773 suggests that we cannot reject the null hypothesis that the risk of death from any cause is not associated with serum LDL levels for subjects with a baseline LDL of 130-160 mg/dL.

*LDL 160 – 190 mg/dL*

 From a proportional hazards regression analysis, we estimate that the instantaneous risk of death is 2.91% lower (hazard ratio 0.9709) for every 1 mg/dL increase in serum LDL among individuals who belong to the group with LDL between 160-190 mg/dL). For every 10 mg/dL increase in serum LDL, there is a decrease in instantaneous risk of death of 25.57% (hazard ratio 0.7443) for subjects with a baseline serum LDL 160-190 mg/dL. Based on a 95% confidence interval, this observed hazard ratio would not be unusual if the true instantaneous risk of death was between 48.29% lower to 14.69% higher for every 10 mg/dL increase in baseline serum LDL. A two-sided p-value of 0.181 suggests that we cannot reject the null hypothesis that the risk of death from any cause is not associated with serum LDL levels for subjects with a baseline LDL of 160-190 mg/dL.

*LDL >190 mg/dL*

 From a proportional hazards regression analysis, we estimate that the instantaneous risk of death is 2.88% higher (hazard ratio 1.0288) for every 1 mg/dL increase in serum LDL among individuals who belong to the group with LDL > 190mg/dL. For every 10 mg/dL increase in serum LDL, there is an increase in instantaneous risk of death of 32.83% (hazard ratio 0.1.3283) for subjects with a baseline serum LDL >190 mg/dL. Based on a 95% confidence interval, this observed hazard ratio would not be unusual if the true instantaneous risk of death was between 19.06% lower to 117.90% higher for every 10 mg/dL increase in baseline serum LDL. A two-sided p-value of 0.261 suggests that we cannot reject the null hypothesis that the risk of death from any cause is not associated with serum LDL levels for subjects with a baseline LDL of >190 mg/dL.

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

Instantaneous risk of death = 0.9789 (if LDL <70mg/dL) + 0.9797 (if LDL 70-100mg/dL) + 0.9977 (if LDL 100-130mg/dL) + 1.0036(if LDL 130-160mg/dL) + 0.9709(if LDL 160-190mg/dL) + 1.0288(if LDL>190mg/dL)

 The slopes correspond to the hazard ratio within each LDL group for each increase in 1mg/dL in baseline serum LDL. Because these are dummy variables, the other terms will drop out as they have values of zero.

* 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 will perform a scatter plot of each hazard regression model with Lowess fitted line to compare the fit of each model. As seen in Figure 4, the data show a possibly u-shaped relationship where the hazard ratio decreases as LDL increases, up to a point when LDL ~190 and the curve starts to uptrend. When LDL is a continuous variable (Figure 3), we see a curvilinear relationship. This model using linear splines appears to be a better fit and also has a chi-square p-value of <0.0001.



Figure 4

* 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.
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?*

The regression strategies from Homeworks 4 and 5 allow us to use censor data. They also allow us to estimate the risk of death based on LDL levels which goes beyond estimating LDL levels based on a binary variable of surviving to 5 years or not surviving to 5 years. Risk is a statistic that is more easily communicated with the public and it can also express the magnitude of how changes in LDL levels can positively or negatively affect mortality.

* 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.

Like the models from Homework 4, these fitted values display a trend that is predominantly downward with higher LDL. Similarly, the greatest variability tends to occur when LDL is low (<70mg/dL) and at levels above 200mg/dL. Also, these models suggest that the relationship between instantaneous risk and serum LDL is likely curvilinear. Although we still cannot determine if the curve is U-shaped because of our sample size, two models in Homework 5 suggested a possible U-shaped trend compared to only one from Homework 4. Also, the y axis is more condensed in the models from Homework 5, suggesting an attenuated risk effect for every unit increase in LDL.



* 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 the flexible modeling of predictors used in this homework because I suspect a nonlinear, potentially complex fit. I assume, as our data has shown, that the trend will be to decreasing hazard ratios as serum LDL increases for values between 70-200mg/dL, but that this trend could reverse at extreme values of LDL, both high and low. I prefer either spline model over the dummy variable model because it allows us to maintain clinically useful LDL groups (based on Mayo Clinic specifications), but without losing the power and degrees of freedom by lumping the values into a binary variable rather than keeping LDL as a “hybrid” continuous variable.