**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2014

**Homework #5**

February 3, 2014

67/77

* 1. *Saturation:* Yes this model is saturated; we have the same number of parameters (4) as we do groups included in the analysis (the 4 race groups).

*Methods:* Robust logistic regression was performed using dummy variables to code for the four races: white (n= 572), black (n= 104), Asian (n= 47), and other (n=12). Dummy variables were encoded so that subjects with the race characteristic of the variable of interest were coded as 1, and all other races were coded as 0. 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 using the approximate normal distribution for logistic regression parameter estimates.

*Inference:* This study provided data on diabetes diagnosis information and race information for 735 individuals. The prevalence of a diagnosis of diabetes is used as a proxy for determining the actual prevalence of diabetes within the study population, as stratified by subject’s race. The odds ratios indicating the odds of a diabetes diagnosis for one race versus the referent race (in this case whites) are stated as follows:

* White subjects had an odds of diabetes diagnosis of 0.1085. These odds would not be unusual if the true odds for white subjects was anywhere between 0.08236 and 0.1430 (p <0.001). Note that in this case the p-value is not informative as it tells us that yes, it is significant that a constant value is different from 0 (because the constant represents the odds ratio when all other covariates are 0, and thus any constant value that has a value will be different from 0).
* Black subjects had an odds ratio for diabetes diagnosis of 1.929, thus exhibiting an elevated odds of diabetes diagnosis in comparison to white subjects. This odds ratio would not be unusual if the true odds ratio for black subjects to white subjects was anywhere between 1.0815 and 3.4391 (p = 0.026). As this observation is statistically significant, we can reject the null hypothesis that there is no difference in the odds of a diabetes diagnosis between black individuals and white individuals.
* Asian subjects had an odds ratio for diabetes diagnosis in relation to white subjects of 0.6282, demonstrating a protective effect of being Asian. This odds ratio would not be unusual if the true odds ratio for Asian subjects to white subjects was anywhere between 0.1888 and 2.091 (p = 0.449). Given the lack of statistical significance of this odds ratio (likely in part due to the statistical imprecision associated with such a small sample size), we fail to reject the null hypothesis that there is no difference in the odds of a diabetes diagnosis between Asian individuals and white subjects.
* Subjects of other races had an odds ratio for diabetes diagnosis in relation to white subjects of 1.843. This odds ratio would not be unusual if the true odds ratio for other race subjects compared to whites was anywhere between 0.3935 and 8.631 (p = 0.438). Given the lack of statistical significance of this odds ratio (likely due to the statistical imprecision associated with such a small sample size), we fail to reject the null hypothesis that there is no difference in the odds of a diabetes diagnosis between ‘other’ race individuals and white individuals.

The answer is thorough and correct.

10/10

* 1. The constant value represents the odds of a diabetes diagnosis in white subjects. The other parameters represent the ratio of the odds of a diabetes diagnosis within black subjects, Asian subjects, or “other” subjects in relation to the odds of diabetes diagnosis in white subjects.

The interpretations are correct. However, values are not reported (-1).

2/3

* 1. The only p-value that is significant is for the odds ratio of diabetes diagnosis in blacks to whites. While the whites p-value is technically significant, in reality it is non-informative as it tells us that yes it is significant that a constant value is different from 0 (because the constant represents the odds ratio when all other covariates are 0, and thus any constant value that has a value will be different from 0).

Correct.

3/3

* 1. When we fit a logistic regression with blacks as the reference group the odds ratios will change, as will the confidence intervals and p-values. This occurs because we are now comparing the odds of different races to blacks instead of whites, and so you are now dividing the odds of diabetes diagnosis in a group by the odds in blacks instead of the odds in whites. The p-values will also be different because now you are seeing whether the odds of diagnosis in a race group are different from blacks, as opposed to whites. However, it is important to note that the actual odds (as opposed to the odds ratios) do not change.

Correctly identified change of reference group. Acknowledges that p-values will change, but does not indicate understanding that the conclusions will remain from previous model since we only have a reparamaterization (-1).

2/3

* 1. The constant value represents the odds of a diabetes diagnosis in black subjects. The other parameters represent the ratio of the odds of a diabetes diagnosis within white subjects, Asian subjects, or “other” subjects in relation to the odds of diabetes diagnosis in black subjects.

The statements are true and correct, but the values are still necessary. (-1)

2/3

* 1. The only p-value that is significant is for the odds ratio of diabetes diagnosis in whites to blacks. While the blacks p-value is technically significant, in reality it is non-informative as it tells us that yes it is significant that a constant value is different from 0 (because the constant represents the odds ratio when all other covariates are 0, and thus any constant value that has a value will be different from 0).

Correct.

3/3

* 1. The p-values change depending on what you use as your reference group, thus if you go only on the p-values, you would be effectively fishing through your data to look for relationships that are statistically significant, but they would not represent a concrete representation of the risk of diabetes diagnosis in comparison to *all* other individuals.

Answer is correct. It seems author understands arbitrary nature of the reference group.

3/3

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)



|  |  |  |
| --- | --- | --- |
| **LDL Categories** | ***n*** | **5 year survival probability** |
| <70 mg/dL | 22 | 59.1% |
| 70-99 mg/dL | 143 | 83.2% |
| 100-129 mg/dL | 228 | 81.1% |
| 130-159 mg/dL | 225 | 87.1% |
| 160-189 mg/dL | 83 | 88.0% |
| >189mg/dL | 24 | 83.3% |

*Methods:* Serum LDL levels were obtained from 725 study participants, of which 131 were observed to die. The minimum LDL measurement was 11 mg/dL, and the maximum LDL was 247 mg/dL. Serum LDL was fit as a dummy variable with divisions of 0-69 mg/dL, 70-99 mg/dL, 100-129 mg/dL, 130-159 mg/dL, 160-189 mg/dL, and 190-249 mg/dL. Cox proportional hazards regression was performed to estimate the instantaneous hazard of death for individuals falling within the above delineated serum LDL concentrations. 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 categorized described above. 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 at the time of study accrual were omitted from the analysis.

*Inference:* Cox proportional hazards regression was performed to explore the instantaneous risk of death from any cause for individuals falling in 6 different categories of serum LDL measurements (see methods above for the category delineations). The hazard ratios for each serum LDL group represent the hazard experienced by individuals in that serum LDL group compared to the baseline level of hazard (hazard for individuals with a serum LDL of 0). Lower and upper 95% CI bounds represent the boundaries on a range of values for which our data would not be unusual if the true hazard ratio were to fall within those bounds. The p-value demonstrates the statistical significance of the inference.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **LDL Level** | **Hazard Ratio** | **Lower 95% CI bound** | **Upper 95% CI bound** | **p-value** |
| 70-99 mg/dL | 0.3980 | 0.1933 | 0.8197 | 0.012 |
| 100-129 mg/dL | 0.3926 | 0.1975 | 0.7804 | 0.008 |
| 130-159 mg/dL | 0.2939 | 0.1451 | 0.5953 | 0.001 |
| 160-189 mg/dL | 0.2565 | 0.1089 | 0.6042 | 0.002 |
| >189mg/dL | 0.3167 | 0.0993 | 1.0107 | 0.052 |

Quite the professional presentation of the results. It is fit for a publication.

10/10

* 1. Cox proportional hazards models do not compute an intercept. Rather, the parameters from the output are hazard ratios that represent the ratio of hazards between two groups that differ in their LDL concentration by 30 units. The baseline hazard pertains to the hazard experienced by a group with an LDL concentration of 0mg/dL. The relationship can then be described as Haz.Ratio= (baseline hazard when LDL=0)^(change in LDL level).

Summary statements are not the key in this problem. Itemized interpretation of parameter estimates are preferred. (-3)

2/5

* 1. I would run a likelihood-ratio test to test the assumption that the simple proportional hazards model of LDL treated as a continuous variable is nested in the categorized LDL model. The results of the likelihood-ratio test (p = 0.4776) demonstrate that there is no evidence that the categorized model fits the data significantly better than when LDL is kept continuous.

Although this answer is not in the key, the likelihood test should give exactly the result desired.

5/5

* 1. Fitted values represent the mean hazard ratios relative to the hazard when LDL is 160 mg/dL for 6 different categories of LDL serum concentrations. See table below.

|  |  |  |
| --- | --- | --- |
| **LDL Categories** | ***n*** | **Hazard Ratio (relative to LDL=160 mg/dL)** |
| <70 mg/dL | 22 | 2.1678 |
| 70-99 mg/dL | 143 | 1.7239 |
| 100-129 mg/dL | 228 | 1.4007 |
| 130-159 mg/dL | 225 | 1.1386 |
| 160-189 mg/dL | 83 | 0.9152 |
| >189mg/dL | 24 | 0.7025 |

Good for you for recognizing that each category only has 1 value.

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
	1. *Descriptive Statistics*



|  |  |  |
| --- | --- | --- |
| **LDL Categories** | ***n*** | **5 year survival probability** |
| <70 mg/dL | 22 | 59.1% |
| 70-99 mg/dL | 143 | 83.2% |
| 100-129 mg/dL | 228 | 81.1% |
| 130-159 mg/dL | 225 | 87.1% |
| 160-189 mg/dL | 83 | 88.0% |
| >189mg/dL | 24 | 83.3% |

*Methods:* Serum LDL levels were obtained from 725 study participants, of which 131 were observed to die. The minimum LDL measurement was 11 mg/dL, and the maximum LDL was 247 mg/dL. Serum LDL was fit using linear splines for LDL divisions of 0-69 mg/dL, 70-99 mg/dL, 100-129 mg/dL, 130-159 mg/dL, 160-189 mg/dL, and 190-249 mg/dL. Cox proportional hazards regression was performed on the spline fit LDL variable to explore the association between serum LDL and the instantaneous risk of death from any cause. 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 at the time of study accrual were omitted from the analysis.

*Inference:* Cox proportional hazards regression was performed on splines fit LDL to explore the instantaneous risk of death from any cause for individuals falling in 6 different categories of serum LDL measurements (see methods above for the category delineations). The hazard ratios for each serum LDL group represent the hazard experienced by individuals in that serum LDL group compared to the baseline level of hazard (hazard for individuals with a serum LDL of 0). Lower and upper 95% CI bounds represent the boundaries on a range of values for which our data would not be unusual if the true hazard ratio were to fall within those bounds. The p-value demonstrates the statistical significance of the inference.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **LDL Level** | **Hazard Ratio** | **Lower 95% CI bound** | **Upper 95% CI bound** | **p-value** |
| <70 mg/dL | 0.9781 | 0.9489 | 1.0082 | 0.152 |
| 70-99 mg/dL | 0.9797 | 0.9537 | 1.0065 | 0.136 |
| 100-129 mg/dL | 0.9977 | 0.9755 | 1.0204 | 0.842 |
| 130-159 mg/dL | 1.0036 | 0.9783 | 1.0295 | 0.782 |
| 160-189 mg/dL | 0.9709 | 0.9282 | 1.0156 | 0.198 |
| >189mg/dL | 1.0288 | 0.9778 | 1.0825 | 0.274 |

Correct and clear.

10/10

* 1. Cox proportional hazards models do not compute an intercept. Rather, the parameters from the output are hazard ratios that represent the ratio of hazards between two groups that differ in their LDL concentration by 30 units. The baseline hazard pertains to the hazard experienced by a group with an LDL concentration of 0mg/dL. The relationship can then be described as Haz.Ratio= (baseline hazard when LDL=0)^(change in LDL level).

Again, summary statements are not the key here. The figures are needed. It’s a repeat error so I will take less points off this time. (-2)

3/5

* 1. I would run a likelihood-ratio test to test the assumption that the simple proportional hazards model of LDL treated as a continuous variable is nested in the categorized LDL model. The results of the likelihood-ratio test (p = 0.3730) demonstrate that there is no evidence that the splines model fit the data significantly better than when LDL is kept continuous.

This approach should work similarly to the multiple partial F test.

5/5

* 1. This was performed in stata.
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. The regression strategies in this homework allowed greater flexibility in terms of what you can make your baseline of interest (as in number 1 of this homework where you can use the regression to compare all other groups to a reference group of interest), and they also allow you to have the ability to fit a regression line (such as with splines) that fits the data more closely than a line that much follow a linear or a quadratic function.

Although not explicitly stated, author is pointing towards precision and advantage of response/predictor of interest.

3/3

* 1. The fitted models from homework 4 are constrained by the fact that their trend needs to take on the shape dictated by the model (for instance, the quadratic modeling requires the curve to be u-shaped). In this homework, the models are more flexible, so the splines fit appears more like a lowess curve of the data, and the dummy variable fit appears as a step-wise function (which is not a particularly good approximation of the data). In addition, because LDL was categorized in the models in this homework, there are fewer fitted values, and the curves are not as continuous as the fitted values in homework 4.

The description is on point but there is no graph. (-2)

1/3

* 1. *A priori* I would not expect this dataset to have a complex non-linear fit, therefore I don’t think that it would be necessary to use flexible models to explore this association. Thus I think that *a priori* the best method to use would be to log transform LDL (since this is a multiplicative model) and use cox proportional hazards regression to explore the relationship between log transformed serum LDL concentrations and the instantaneous risk of death from any cause. The logarithmic scale for LDL is more interpretable than other transformations of LDL, and it appears to be a more biologically plausible transformation for the variable.

Excellent.

3/3