Homework 5

1. Regression analysis evaluating an association between prevalence of diabetes and race by comparing the odds of a diabetes diagnosis across groups.
2. The logistic regression model using whites as a reference group is a saturated model since there are 4 parameters and 4 groups. Our analysis is based on a logistic regression model evaluating the odds of death using Whites as a baseline group with respect to Asians, Blacks, and others. We do not use the robust estimator since the estimate obtained by the robust estimator is not much different from our default, which computes our standard errors and confidence intervals using the likelihood test.

Based on the p-value of 0.1096, we cannot reject the null hypothesis of no association between diabetes and race (or that the two groups have an odds ratio of diabetes of 1). The Black group has odds of diabetes 1.9286 times higher than that of Whites, with the 95% CI suggesting that what we observe would not be surprising if blacks had anywhere from 1.0819 times to 3.4377 times the odds of diabetes compared to the white group. The Asian group has odds of diabetes 0.6282 times as large as that of whites, with the 95% CI suggesting that what we observe would not be surprising if Asians had anywhere from 0.1889 times to 2.0892 times the odds of diabetes compared to the White group. The other races group has odds of diabetes 1.8429 times higher than that of Whites, with the 95% CI suggesting that what we observe would not be surprising if the other races had anywhere from 0.3939 times to 8.6222 times the odds of diabetes compared to the white group.

1. The model is log (p/1-p) = b0+b1\*X1+b2\*X2+b3\*X3. The intercept (b0) represents the odds of diabetes for the White group. The parameters b1, b2, and b3 represent the odds ratios of diabetes of Blacks, Asians, and Others groups (respectively) compared to the White group.
2. If we were to ignore issues related to multiple comparisons, we would conclude that the p-value is only significant for the Black group compared to the White group. Based on the significant p-value of 0.026 for the Black group, we may conclude that the test is significant and that we can reject the null hypothesis that the odds ratio of diabetes between Blacks and Whites is 1.
3. Fitting a logistic regression model using blacks as a reference group, our odds ratios and confidence intervals will change since now our baseline group is the Blacks. Now our inferences are reciprocals of what we obtained in part (a). However, if we interpret our problem using the logit model instead where the odds of diabetes is reported, then our formal inferences would not change. Thus, our regression model is merely a reparameterization of the model in part (a).
4. The intercept (b0) represents the odds of diabetes for the Black group. The parameters b1, b2, and b3 represent the odds ratios of diabetes of Whites, Asians, and Others groups (respectively) compared to the Black group.
5. If we ignore the issues related to multiple comparisons, we would conclude that the p-value is only significant for the White group compared to the Black group. Based on the significant p-value of 0.026 for the White group, we may conclude that the test is significant and that we can reject the null hypothesis that the odds ratio of diabetes between Blacks and Whites is 1.
6. We should not ignore any of the p-values for individual regression parameters from a dummy variable regression, and we should include our complete data in the results. It is possible that our results for significance will change if we decide to reject parameters from the regression model based on their insignificance and re-run the analysis using just the other variables.
7. Regression analysis evaluating an association between mortality and LDL comparing instantaneous risk of death over the entire period of observation.
8. We perform a proportional hazards regression modeling LDL as a dummy variable. The 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 the Wald statistics based on the Huber-White sandwich estimator using the robust option.

Descriptive statistics for the censoring distribution includes the maximum observed censoring times and the Kaplan-Meier estimates of the 10th, 50th (median), 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 survival curve.

Descriptive statistics for censoring distribution:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **death = 0** |  |  |  | **death=1** |  |  |
| ldlCTG | N | max |  | ldlCTG | N | max |
| 0 | 12 | 5.7550 |  | 0 | 10 | 5.1663 |
| 70 | 115 | 5.8782 |  | 70 | 28 | 5.4949 |
| 100 | 184 | 5.8836 |  | 100 | 44 | 5.3580 |
| 130 | 191 | 5.9083 |  | 130 | 34 | 5.5359 |
| 160 | 72 | 5.9055 |  | 160 | 11 | 5.3881 |
| 190 | 20 | 5.9083 |  | 190 | 4 | 4.9446 |
| Total | 594 | 5.9083 |  | Total | 131 | 5.5359 |



|  |
| --- |
| **Censoring Times** |
| 10th | 5.0294 |
| 50th | 5.1855 |
| 90th | 5.7769 |
| rest. mean | 5.333 |

Descriptive statistics for serum LDL levels included the number of cases with missing data, as well as the minimum, maximum, mean, standard deviation, and the quantiles for the cases with available data. The serum LDL levels were categorized according to the directions given in the homework, which were: 0-70 mg/dL, 70-99 mg/dL, 100-129 mg/dL, 130-159 mg/dL, 160-189 mg/dL, 190-249 mg/dL, and above 250 mg/dL.

Descriptive statistics for serum LDL:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| variable | N | mean | sd | min | p25 | p50 | p75 | max |
| ldl | 725 | 125.8028 | 33.6020 | 11 | 102 | 125 | 147 | 247 |

Kaplan-Meier survival estimates:



Based on the proportional hazards regression model computed using methods mentioned above, we notice that data was available on 725 subjects having a mean serum LDL of 125.8028 mg/dL and a standard deviation of 33.6020 mg/dL. The range of LDL measures is 11 mg/dL to 247 mg/dL. During an average of 5.33 of observation, 131 of the subjects were observed to die. Modeling LDL as a dummy variable with the categories defined according to directions given in the homework, we use the group with 0-70 mg/dL as the baseline and obtain hazard ratios for the other groups relative to the baseline group. We estimate that the instantaneous risk of death is a relative 60.20% lower (hazard ratio 0.3980), 60.74% lower (hazard ratio 0.3926), 70.61% lower (hazards ratio 0.2939), 74.35% lower (hazards ratio 0.2565), and 68.33% lower (hazards ratio 0.3167) for the groups with LDL measures of 70 to 99 mg/dL, 100 to 129 mg/dL, 130 to 159 mg/dL, 160 to 189 mg/dL, and above 190 mg/dL, respectively, relative to the baseline group which had LDL measures of 0 to 69 mg/dL. A 95% CI suggests that our observed ratios are not unusual if the true instantaneous risks of death were anywhere from 79.74% to 21.80% lower, 79.29% to 25.58% lower, 84.79% to 43.22% lower, 88.65% to 42.01% lower, and 89.86% to 1.08% lower for the groups with LDL measures of 70 to 99 mg/dL, 100 to 129 mg/dL, 130 to 159 mg/dL, 160 to 189 mg/dL, and above 190 mg/dL, respectively, relative to the baseline group which had LDL measures of 0 to 69 mg/dL. Based on our p-value of 0.0087, we can reject the null hypothesis of no association between all-cause mortality and LDL levels in favor of studies promoting association between the two.

1. The intercept refers to the baseline hazard, but generally is not looked at during the experiment. Our slope is obtained by exponentiating the hazards ratio reported by Stata, where a group with 1 mg/dL higher serum LDL has event rate higher by a factor of the reported hazards ratio (our parameter). Equivalently, to use our hazards ratio parameter to compute the instantaneous event rate for a group with 10 units higher measurement of serum LDL, we can exponentiate the reported hazards ratio by 10.
2. In order to assess whether the regression model provides a better fit than the model that uses only a continuous linear term for LDL, we can perform a chi squared test for the goodness of fit. We obtain an insignificant p-value of 0.3988, which means that we cannot reject the null hypothesis that there is a linear association between LDL and mortality rate. We can also infer that there is no nonlinearity in the data.
3. Relative hazards ratios variable generated.
4. Regression analysis evaluating an association between mortality and LDL comparing instantaneous risk of death over the entire period of observation using splines.
5. We perform a proportional hazards regression using linear splines of LDL. Linear splines were created using the directions from homework, which were to cut the data at 70 mg/dL, 100 mg/dL, 130 mg/dL, 160 mg/dL, and 190 mg/dL levels of LDL. This creates 6 partitions of the data, within which we perform individual proportional hazards regression. The estimated hazards ratio is computed within each group, with the null hypothesis being that there is no association between all-cause mortality rate and LDL within each stratum. We also use the robust estimator since there is no reason to assume equal variance between the groups. The standard error and confidence interval is computed using Wald-based tests.

For the descriptive statistics of all-cause mortality, which is censored, and for the descriptive statistics of high serum LDL, we can infer to problem 2 since we are merely modeling the two variables differently.

Based on the highly significant two-sided p-value (P<0.0001), we can reject the null hypothesis of no association between all-cause mortality rate and LDL. The estimated hazard of death decreases by 2.2% (95% CI -4.0% to -0.4%), 2.0% (95% CI -4.7% to +0.7%), 0.3% (95% CI -2.4% to +1.9%), and 2.9% (95% CI -7.0% to +1.4%) for the groups with 0 to 69 mg/dL, 70 to 99 mg/dL, 100 to 129 mg/dL, and 160-189 mg/dL, respectively. The estimated hazard of death increases by 0.4% (95% CI -2.1% to +2.8%) and 2.9% (95% CI -2.1% to +8.1%), respectively, for the groups with 130 to 159 mg/dL and above 190 mg/dL.

1. In each group, the intercept is usually not of interest although if we were to estimate it, it would refer to the hazard at baseline for each group. Our parameter in each group is the hazard ratio, with the estimated hazard ratio being the ratio of hazards between two groups differing by 1 mg/dL of LDL. This ratio is variable depending on which group we are looking at.
2. In order to assess whether the regression model provides a better fit than the model that uses only a continuous linear term for LDL, we can perform a chi squared test for the goodness of fit. We obtain a insignificant p-value of 0.0788, which means that we cannot reject the null hypothesis that there is a linear association between LDL and mortality rate. We can also infer that there is no nonlinearity in the data.
3. Relative hazards ratios variable generated.
4. Advantages/disadvantages of various statistical analysis strategies
5. Homework 4 and 5 focuses on proportional hazards regression measuring instantaneous risk of death, and homework 1-3 focus on a variety of regression methods modeling the variables LDL and mortality categorically and continuously. The strategies used in homework 4 and 5 allow us to use censored data using Kaplan-Meier estimates, which lets us use the complete information contained in the data. The dummy-variable and spline model in homework 5 also allow for more specialized and flexible analysis of the data, and allow us to capture more information about the data itself, therefore having to borrow less information from the model we construct.
6. There is an overall downward trend in relative hazard ratios, with all the values centered at 160 mg/dL. The differences between the fitted values from the three models in homework 4 and the two models fit in problems 2 and 3 of this homework are that the models from homework 4 are smoother, while the dummy fit model in this homework is a jump function due to the categorical nature of the model, and the spline fit model in this homework is more specialized to look at different sections of the data. The models in homework 4 assumed an equal distribution of death as a function of LDL, but the models in homework 5 did not.



1. A priori, I would prefer the splines model for proportional hazards regression. This is because we are able to look at continuous patterns of the data partitioned by different strata of LDL. This allows for a more flexible model taking into account the by-strata difference of the effect of LDL measure on all-cause mortality. The dummy variable model is not as good because we use a categorization of LDL, and categorical variable is not a good way to analyze a continuous model. Other continuous models are also good, but in order to take into account censoring and potential by-strata difference in distribution and variance (heteroscedasticity), we would prefer the proportional hazards regression using the splines model.