**Homework #1**

January 6, 2014

1. The observations of time to death in this data are subject to (right) censoring. Nevertheless, problems 2 – 6 ask you to dichotomize the time to death according to death within 5 years of study enrolment or death after 5 years. Why is this valid? Provide descriptive statistics that support your answer.

When the data is dichotomized according to 5-year survival, the question is simply whether or not the subject was alive at 5 years after study enrollment and it is unimportant if they died after 5 years. If a subject is lost to follow-up prior to the 5 years, however, their data would be censored since we do not know whether they died before or after 5 years. To support the validity of dichotomization of this censored data, the proportion of subjects dead within 5 years was calculated for all subject data. This proportion will not change regardless of the time and number of deaths that occur after the 5-year point. Additionally, the median survival will not change with right-censored data that occurs after 5 years. Since the median is simply a measure of time at which half of the population has reached the endpoint (death), the number of deaths occurring after half the population has died will not affect the median.

1. Provide a suitable descriptive statistical analysis for selected variables in this dataset as might be presented in Table 1 of a manuscript exploring the association between serum LDL and 5 year all-cause mortality in the medical literature. In attention to the two variables of primary interest, you may restrict attention to age, sex, weight, smoking history, and prior history of cardiovascular disease (coronary heart disease (CHD), congestive heart failure (CHF), and stroke.

Method: The data was dichotomized into four groups based on LDL measurements for ideal (<130 mg/dL), borderline (130-159 mg/dL), high (160-189 mg/dL), and very high (≥ 190 mg/dL). Descriptive statistics were then generated to summarize the demographic variables of interest.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Ideal LDL(< 130 mg/dL) | Borderline LDL(130-159 mg/dL) | High LDL(160-189 mg/dL) | Very High LDL(≥ 190 mg/dL) |
| Age (yr), mean (SD) | 74.7 (5.3) | 74.2 (5.6) | 74.6 (5.7) | 75.5 (6.1) |
| Male (%) | 55.5 | 43.1 | 48.2 | 32.4 |
| Weight (lb), mean (SD) | 159.9 (29.9) | 158.4 (32.3) | 165.1 (32.9) | 158.1 (22.9) |
| History of Smoking (%) Pack years, mean (SD) | 56.019.8 (26.9) | 57.320.0 (28.8) | 57.819.3 (23.8) | 50.014.9 (25.5) |
| History of Cardiovascular Disease (%) Congestive heart failure Coronary heart disease Stroke | 6.634.622.1 | 4.931.121.8 | 2.428.930.1 | 5.947.038.2 |
| Survival time (days),  Mean Median | 1778 (403.2)1869 | 1832 (397.8)1887 | 1858 (282.6)1875 | 1792 (541.5)1916 |

1. Perform a statistical analysis evaluating an association between serum LDL and 5 year all-cause mortality by comparing mean LDL values across groups defined by vital status at 5 years.

Method: Patients were divided into two groups: those who survived 5 years, and those who did not. Mean LDL values were generated for each group, and the means were compared using a two-sided two-sample t-test with α = 0.05 and unequal variances.

Inference: The mean LDL measurement for patients who were alive at 5 years was 127.4 mg/dL (95% CI 124.7 - 130.0 mg/dL), and the mean LDL for patients who did not survive 5 years was 118.6 mg/dL (95% CI 112.4 -124.8 mg/dL). The null hypothesis is that the mean LDL values between groups are equal. Analysis by a two-sample t-test showed that the difference in means is 8.75 mg/dL (95% CI 2.0 – 15.5 mg/dL) with a two-sided p-value of 0.0113. This p-value is less than the predetermined threshold of 0.05, providing evidence to reject the null hypothesis, and suggesting the difference in mean LDL is greater than that predicted under the null hypothesis.

1. Perform a statistical analysis evaluating an association between serum LDL and 5 year all-cause mortality by comparing geometric mean LDL values across groups defined by vital status at 5 years.

Method: Patients were grouped as defined in question (3) above. The geometric mean LDL values were generated for each group, and were compared using a two-sided two-sample t-test with α = 0.05 and unequal variances.

Inference: The null hypothesis states that the geometric mean LDL values between groups are equal. Analysis by a two-sample t-test gave a two-sided p-value of 0.0075. This p-value is less than the predetermined threshold of 0.05, providing evidence to reject the null hypothesis, and suggesting the difference in mean LDL is greater than that predicted under the null hypothesis.

1. Perform a statistical analysis evaluating an association between serum LDL and 5 year all-cause mortality by comparing the probability of death within 5 years across groups defined by whether the subjects have high serum LDL (“high” = LDL > 160 mg/dL).

Method: Patients were grouped according to LDL levels into “high” (LDL > 160 mg/dL) and “low” (LDL < 160 mg/dL). The proportion of patients in each group who died prior to reaching the 5-year follow-up point was calculated for each group, a chi-squared analysis was conducted, and the relative risk was calculated to estimate the relative probability of death in each group. This allows a comparison to identify evidence of whether death is dependent upon LDL measurements.

Inference: The null hypothesis states that the probability of death will not vary based on LDL values, and thus the risk ratio will be equal to one. A Chi-squared test gave a p-value of 0.2746, which is greater than the predetermined threshold of 0.05, providing insufficient evidence to reject the null hypothesis. The risk of death in the group with “low” LDL is 18.8%, and the risk among those with “high” LDL is 14.5%. The risk ratio is 1.29 (95% CI 0.808 – 2.06). Since this confidence interval contains 1, this supports the conclusion that there is no evidence of an association between probability of death and LDL levels.

1. Perform a statistical analysis evaluating an association between serum LDL and 5 year all-cause mortality by comparing the odds of death within 5 years across groups defined by whether the subjects have high serum LDL (“high” = LDL > 160 mg/dL).

Method: Patients were grouped according to LDL levels into “high” (LDL > 160 mg/dL) and “low” (LDL < 160 mg/dL). Odds is defined as the probability of death divided by the probability of survival. The proportion of patients in each group who died prior to reaching the 5-year follow-up point was calculated for each group, a chi-squared analysis was conducted, and the odds ratio was calculated to estimate the relative probability of death in each group. This allows a comparison to identify evidence of whether death is dependent upon LDL measurements.

Inference: The null hypothesis states that the probability of death will not vary based on LDL values, and thus the odds of death will be equal between groups. A Chi-squared test gave a p-value of 0.2746, which is greater than the predetermined threshold of 0.05, providing insufficient evidence to reject the null hypothesis. The odds of death in the group with “low” LDL is 23.1%, and the risk among those with “high” LDL is 17.0%. The odds ratio is 1.36 (95% CI 0.782 – 2.36). Since this confidence interval contains 1, this supports the conclusion that there is no evidence of an association between probability of death and LDL levels.

1. Perform a statistical analysis evaluating an association between serum LDL and all-cause mortality over the entire period of observation of these subjects by comparing the instantaneous risk of death across groups defined by whether the subjects have high serum LDL (“high” = LDL > 160 mg/dL).

Method: Patients were grouped according to LDL levels into “high” (LDL > 160 mg/dL) and “low” (LDL < 160 mg/dL). Using linear regression, the data were analyzed using LDL values as the ‘Y’ variable, and observation time, which records the number of days to death or the end of the study, as the ‘X’ variable. This will identify an association between LDL and death throughout the study period.

Inference: The null hypothesis states that the probability of death will not vary based on LDL values. By linear regression, the p-value for an association between LDL and time to death was 0.107 for the “low” LDL group and 0.934 for the “high” LDL group. This suggests there is insufficient evidence of an association between death and LDL levels, and thus we fail to reject the null hypothesis.

1. Supposing I had not been so redundant (in a scientifically inappropriate manner) and so prescriptive about methods of detecting an association, what analysis would you have preferred *a priori* in order to answer the question about an association between mortality and serum LDL? Why?

I would have chosen to do linear regression since LDL is a continuous variable and death at 5-years is binary, regression provides the most robust analysis to detect an association between death and LDL levels.