**1174**

**Biost 515 (Winter 2014)**

**Instructor: Scott Emerson**

**Homework 1**

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 enrollment or death after 5 years. Why is this valid? Provide descriptive statistics that support your answer.

**ANSWER:** Despite (right) censoring, the data from this study can be dichotomized according to participants’ time to death within five years of study enrollment or after five years of study enrollment. The cohort study included 735 participants, 602 of which were alive at the end of the study period and therefore have right censored observations of time to death, and 133 of which were observed to die while on the study and therefore have uncensored times to death. The minimum censored observation of time to death is 1827 days, or 5.002 years. Therefore, dichotomization at five years of follow-up does not needlessly embed left censoring into the data, and more importantly, provides us with a subset of data which is completely uncensored (the subset being those observations with time to death less than or equal to five years). In questions 2-6, we focus our attention towards 5 year all-cause mortality. By our methods of dichotomization, we have eliminated all censored observations of time to death in the subset of interest and can thus avoid more cumbersome survival analysis techniques.

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.

**ANSWER:** Table 1 presents baseline characteristics for selected variables of the 735 participants in the study population. As shown, the results are first dichotomized according to Mayo Clinic recommendations for risk of cardiovascular disease given various levels of serum LDL (low density lipoprotein), and then by the overall study population. While the primary variables of interest are serum LDL and 5 year all-cause mortality, attention has also been given to age, sex, weight, history of smoking, and prior history of coronary heart disease (CHD), congestive heart failure (CHF), and any cerebrovascular events.



As shown in Table 1, 61.6% of participants (n = 453) measured serum LDL levels between 100-159 mg/dL. Participants were similar across all serum LDL levels with respect to sex (49.8% male), mean age (74.6 years), and mean weight (159.9 pounds). The proportion of participants who reported a history of smoking was greatest among those with serum LDL between 100-129 mg/dL (59.7%), while the overall average was 56.6%. Altogether, 21.1% of participants had previously been diagnosed with coronary heart disease (CHD), 5.6% with congestive heart failure (CHF), and 13.5% with a cerebrovascular event. Five year all-cause mortality was substantially highest among participants with serum LDL below 70mg/dL (40.9%); while similar mortality rates were recorded for the remainder of participants (12.1% - 18.9%).

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.

**ANSWER:** In order toevaluate an association between serum LDL and 5 year all-cause mortality, this analysis uses a two sample, two sided t-test comparing differences in mean LDL across groups defined by 5 year vitality status. Results were interpreted at an alpha = 0.05 level of significance. The mean serum LDL among participants who were observed to die within five years was 118.7 mg/dL and the mean LDL among those surviving past five years was 127.2 mg/dL. Therefore, the difference in means between the two groups is 8.5 (95% CI: 1.4 - 15.6). The data for this sample population produce strong evidence to demonstrate that such a large difference in mean LDL would be unlikely if there is truly no association between serum LDL and 5 year all-cause mortality (p = 0.0186).

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.

**ANSWER:** This analysis, which seeks the same associations as above, utilizes a two sample, two sided t-test comparing differences in serum LDL geometric means across groups defined by 5 year vitality status. Results are interpreted at an alpha = 0.05 level of significance. The difference in geometric means between the two groups was determined to be 1.09 (95% CI: 1.02 – 1.18). The data for this sample population produce strong evidence to show that, if there is truly no association between serum LDL and 5 year all-cause mortality, our results for differences in serum LDL geometric means would be unlikely (p = 0.0128).

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).

**ANSWER:** This analysis evaluates an association between serum LDL and 5 year all-cause mortality by calculating the difference in risk of death between participants with high LDL (> 160 mg/dL) verses those with low LDL (<160 mg/dL). Generalized linear modeling was utilized to calculate this risk difference and significance was determined at alpha = 0.05. As modeled, the risk of death for participants with low serum LDL is 0.04 more than participants with high serum LDL (95% CI: -0.03 – 0.10). Unfortunately, the data for this study lack the precision necessary to demonstrate that a risk difference of this size between groups is unlikely in the absence of a true association (p = 0.277).

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).

**ANSWER:** This analysis evaluates an association between serum LDL and 5 year all-cause mortality by calculating the odds ratio for mortality between participants with high LDL (> 160 mg/dL) verses those with low LDL (<160 mg/dL). Generalized linear modeling was utilized to calculate the odds ratio and significance was determined at alpha = 0.05. As calculated, the odds of all-cause mortality within 5 years were 1.4 times larger for participants with low LDL than participants with high LDL (95% CI: 0.7 – 2.5). Unfortunately, the data for this study are statistically insignificant, and therefore fail to indicate that these results would be unlikely in the absence of a true association (p = 0.315).

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).

**ANSWER:** In order to compare the instantaneous risk of death across groups defined by high serum LDL (> 160 mg/dL) and low serum LDL (<160 mg/dL), this analysis utilized Kaplan-Meier survival estimates to determine instantaneous survival probabilities and hazard ratio estimates based on a logrank test and cox proportional hazards model, respectively. Statistical significance was determine at a level of alpha = 0.05. Under the logrank test, we fail to reject the notion of equal survival probabilities between the high and low LDL groups (p = 0.225). Moreover, under proportional hazards regression, for every 0.7 deaths among high LDL participants, there is 1 death among low LDL participants (95% CI: 0.4 – 1.2). Unfortunately, if we assume no association between serum LDL and mortality, this study lacks the precision necessary to demonstrate these results are unlikely.

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?

**ANSWER:** Had the methods of detecting an association between serum LDL and 5 year all-cause mortality not been so redundant and prescriptive, my a priori analysis would have included Kaplan-Meier estimates for survival analysis over the entire period of observation, dichotomizing the data according to high serum LDL levels (> 160 mg/dL) and low serum LDL levels (<160 mg/dL). Specifically, I would have used the logrank test in order to detect a difference in survival probabilities between the two groups. This seems to be a logical analysis for multiple reasons: Kaplan-Meier methods adjust well for censored data; our analysis focuses on how the predictor of interest– serum LDL levels – affects the outcome of interest – mortality, rather than vice versa (as was done in questions 3and 4); and finally, we are avoiding dichotomization of the data, so in some sense, our results are more descriptive of mortality rates overall.