All questions relate to associations between death from any cause and serum low density lipoprotein (LDL) levels in a population of generally healthy elderly subjects in four U.S. communities. This homework uses the subset of information that was collected to examine MRI changes in the brain. The data can be found on the class web page (follow the link to Datasets) in the file labeled mri.txt. Documentation is in the file mri.pdf. The data is in free-field format, and can be read into Stata using the following code in a .do file.

infile ptid mridate age male race weight height packyrs yrsquit alcoh ///

physact chf chd stroke diabetes genhlth ldl alb crt plt sbp aai ///

fev dsst atrophy whgrd numinf volinf obstime death ///

using http://www.emersonstatistics.com/datasets/mri.txt

Note that the first line of the text file contains the variable names, and will thus be converted to missing values. Similarly, there is some missing data recorded as ‘NA’, and those, too, will be converted to missing values. If you do not want to see all the warning messages, you can use the “quietly” prefix. You may want to go ahead and drop the first case using “drop in 1”, because it is just missing values.

Recommendations for risk of cardiovascular disease according to serum LDL (low density lipoprotein) levels are as follows (taken from the Mayo Clinic website):

|  |  |
| --- | --- |
| Below 70 mg/dL | Ideal for people at very high risk of heart disease |
| Below 100 mg/dL | Ideal for people at risk of heart disease |
| 100-129 mg/dL | Near ideal |
| 130-159 mg/dL | Borderline high |
| 160-189 mg/dL | High |
| 190 mg/dL and above | Very high |

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.

**The scientific question addressed in problems 2-6 is interested in an association between serum LDL levels and 5 year all-cause mortality. Because survival is dichotomized to less than 5 years and greater than 5 years the true time of death for participants who live longer than 5 years on the study is irrelevant. The only factor that needs to be known is whether they died within 5 years of starting the study or not. This is a valid approach for this data set because all participants were observed to die during the study or were followed for at least 5 years after enrollment. The minimum observation time for a participant who did not die while on the study was 1,827 days (5 years = 1,825 days).**

|  |  |  |
| --- | --- | --- |
|  | **Minimum Observation Time (days)** | **Maximum Observation Time (days)** |
| **Participant Death During Study (n = 133)** | 68 | 2,022 |
| **Participant Alive at End of Observation (n = 602)** | 1,827 | 2,159 |

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.



**1CHD = Coronary Heart Disease; MI = Myocardial Infarction**

**2CHF = Congestive Heart Failure**

**3CE = Cerebrovascular Event; TIA = Transient Ischemic Attack**

**4 Descriptive statistics were calculated using values for participants with a history of smoking**

**Methods: Three binary variables were created, one indicating a participant’s vital status after 5 years of observation, one indicating whether a participant had ever smoked, and one indicating whether the participant’s serum LDL level was normal (less than 160 mg/dL) or high (160 mg/dL or higher). Descriptive statistics were then calculated for each serum LDL group. The frequency of each observation was calculated for binary and categorical variables; 5 year vital status, gender, prior history of diagnoses, and smoking history were included. Descriptive statistics calculated for continuous variables were mean, standard deviation, minimum, maximum, and quartiles. Serum LDL, age, weight and pack years of smoking were included as continuous variables. Descriptive statistics for pack years of smoking were calculated only using participants with a history of smoking. Participants with no history of smoking were excluded from these calculations as they all had zero pack years and would have skewed the statistics towards zero.**

**Inference: More than four fifths of the observed participants had normal serum LDL, defined as levels less than 160 mg/dL. The average age, weight, gender, medical diagnosis history and smoking history were similar for both groups. 5 year all-cause mortality was slightly higher for participants with serum LDL levels below 160 mg/dL.**

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.

**Methods: A binary variable was created indicating a participant’s vital status after 5 years of observation. The mean serum LDL levels were then compared for vital status groups using a two-sided two-sample t-test that allowed for unequal variance. An alpha level of 0.05 was used and the 95% confidence interval was calculated. The null hypothesis was that there is no difference in mean serum LDL across groups defined by vital status at 5 years, µalive - µdead = 0. The alternative hypothesis is that there is a difference in mean serum LDL across groups defined by vital status at 5 years, µalive - µdead ≠ 0.**

**Inference: The mean serum LDL level for participants who died within 5 years was estimated as 118.70 mg/dL and the mean serum LDL level for participants who survived at least 5 years was estimated as 127.20 mg/dL. The difference in mean serum LDL levels was therefore estimated as being 8.50 mg/dL less for participants who died within 5 years of observation. A 95% confidence interval suggests that this observation is not unusual if the true difference in mean serum LDL for groups defined by vital status at 5 years was between 1.44 and 15.56 mg/dL less for participants who died within 5 years. Because the p-value is less than 0.05 (p = 0.0186) we reject the null hypothesis that there is no difference in mean serum LDL levels across groups defined by vital status at 5 years.**

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.

**Methods: A binary variable was created indicating a participant’s vital status after 5 years of observation. A variable that was the log transformed values of serum LDL levels was also created. The geometric mean serum LDL levels were then compared for vital status groups using a two-sided two-sample t-test that allowed for unequal variance on the log transformed LDL values. The results were then exponentiated to give the point estimate and 95% confidence interval for the difference in geometric means for vital status groups interpretable on the original measurement scale. An alpha level of 0.05 was used. The null hypothesis was that there is no difference in the geometric mean of serum LDL levels across groups defined by vital status at 5 years, GMalive - GMdead = 0. The alternative hypothesis is that there is a difference in the geometric mean of serum LDL levels across groups defined by vital status at 5 years, GMalive - GMdead ≠ 0.**

**Inference: The difference in the geometric mean of serum LDL levels is estimated by 9.65% less for participants who died within 5 years of observation. A 95% confidence interval suggests that this observation is not unusual if the true difference in the geometric mean of serum LDL levels for groups defined by vital status at 5 years was between 2.01 and 17.9%less for participants who died within 5 years. Because the p-value is less than 0.05 (p = 0.0128) we reject the null hypothesis that there is no difference in the geometric mean of serum LDL levels across groups defined by vital status at 5 years.**

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).

**Methods: Two binary variables were created, one indicating a participant’s vital status after 5 years of observation and one indicating whether the participant’s serum LDL level was normal (less than 160 mg/dL) or high (160 mg/dL or higher). The probability of death within 5 years for high and normal serum LDL groups was compared by taking the ratio of the probabilities. An alpha level of 0.05 was used and a 95 % confidence interval was calculated. The null hypothesis was that there is no difference in probability of death across groups defined by serum LDL levels and the ratio of probabilities is equal to one, Phigh/Pnormal = 1.0. The alternative hypothesis was that there is a difference in probability of death across groups defined by serum LDL levels and the ratio of probabilities is not equal to one, Phigh/Pnormal ≠ 1.0.**

**Inference: The probability of death within 5 years was estimated as 0.14 for participants with serum LDL levels greater than or equal to 160 mg/dL and 0.17 for participants with serum LDL levels less than 160 mg/dL. The ratio of these probabilities is estimated by 0.805, indicating a 19.51% decrease in probability of death within 5 years for participants with serum LDL levels of at least 160 mg/dL. A 95% confidence interval suggests that this observation is not unusual if the true ratio for probability of death within 5 years between serum LDL groups was between 0.494 and 1.311. Because the p-value is greater than 0.05 (p = 0.3753) we fail to reject the null hypothesis that there is no difference in probability of death across groups defined by serum 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).

**Methods: Two binary variables were created, one indicating a participant’s vital status after 5 years of observation and one indicating whether the participant’s serum LDL level was normal (less than 160 mg/dL) or high (160 mg/dL or higher). The odds of death within 5 years for high and normal serum LDL groups were compared by taking the ratio of the odds. An alpha level of 0.05 was used and a 95 % confidence interval was calculated using the “Woolf” option. The null hypothesis was that there is no difference in odds of death across groups defined by serum LDL levels and the odds ratio is equal to one, oddshigh/oddsnormal = 1.0. The alternative hypothesis was that there is a difference in odds of death across groups defined by serum LDL levels and the odds ratio is not equal to one,**

**oddshigh/oddsnormal ≠ 1.0.**

**Inference: The ratio for odds of death within 5 years across serum LDL levels is estimated by 0.735, indicating a 26.5% decrease in odds of death within 5 years for participants with serum LDL levels of at least 160 mg/dL. A 95% confidence interval suggests that this observation is not unusual if the true odds ratio was between 0.404 and 1.34. Because the p-value is greater than 0.05 (p = 0.3753) we fail to reject the null hypothesis that there is no difference in odds of death across groups defined by serum 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).

**Methods: A binary variable was created indicating whether the participant’s serum LDL level was normal (less than 160 mg/dL) or high (160 mg/dL or higher). The instantaneous risk of death for each serum LDL group was given by the Kaplan Meier hazard curve, generated using the observation time in years as the time to event and death as the outcome. The log-rank test was used to compare the curves for the groups defined by serum LDL levels. An alpha level of 0.05 was used. The null hypothesis was that there is no difference in the instantaneous risk of death across groups defined by serum LDL levels so the number of observed deaths is equal to the number of expected deaths for each group. The alternative hypothesis was that there is a difference in the instantaneous risk of death across groups defined by serum LDL levels and the observed number of deaths will differ from the expected number of deaths for each group.**

|  |  |
| --- | --- |
|  |  |

**Inference: The observed number of deaths during the study for participants with serum LDL levels below 160 mg/dL was 116 when 111.26 deaths were expected. For participants with serum LDL levels of at least 160 mg/dL there were 17 deaths observed and 21.74 were expected. Because the p-value is greater than 0.05 (p = 0.2664) we fail to reject the null hypothesis that there is no difference in instantaneous risk of death across groups defined by serum LDL levels. This is additionally supported by visual inspection of the figures above, there is a high degree of overlap of the confidence intervals for each curve and the curves cross several times indicating that they are not significantly different.**

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?

**In order to investigate whether an association between serum LDL levels and mortality is present the following analytical plan will be used. The response variable is survival and the predictor of interest is serum LDL. Survival was selected as the response variable because it is of more scientific interest to predict the effect serum LDL has on survival than to predict serum LDL based on survival. In order to obtain scientifically relevant results serum LDL levels will be split into the clinically relevant divisions shown at the beginning of this document. This should ensure that any associations indicating increased or decreased survival within LDL groups that are observed can easily be translated into potential clinical actions. The mean length of survival will then be compared across groups defined by serum LDL levels while adjusting for potential confounders and effect modifiers such as age, gender, weight, medical history, and smoking history. If participants are not lost to follow-up regression analysis can also be used to describe the relationship between serum LDL and survival. If participants are lost to follow-up Kaplan Meir methods will be used to account for censoring of data and survival probabilities will be compared across the groups defined by serum LDL levels.**