1. To address the validity of dichotomizing the time to death according to death within or after 5 years of study enrollment, I prepare a frequency table of the number of censored observations within these time intervals. The table below is a summary of descriptive statistics obtained.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Dead** | | **Total** |
| **Censored observations** | No | Yes |  |
| More than 5 yrs | 602 | 12 | 614 |
| Less than or equal to 5 yrs | 0 | 121 | 121 |

From the table of descriptive statistics, we observe that there were no cases of censoring for all patients that survived for less than or up to five years. Therefore, analyses based on dichotomizing subjects with respect to 5 year survival are valid.

1. The table below provides descriptive statistics within groups defined by low density lipoprotein (ldl) levels (below 160 mg/dl vs 160 mg/dl or above). There are 618 subjects whose ldl was below 160 mg/dl and 110 with ldl greater than or equal to 160 mg/dl. There are 10 ldl missing values. Measurements for smoking history are missing for one subject. Age, sex, and smoking history of the subject don’t appear to have any significant effect on ldl levels of the subject. Patients with high ldl level have a slight trend towards being heavier than subjects with low ldl level. With respect to coronary heart disease (chd), congestive heart failure (chf), and stroke (levels 1 and 2), most of the subjects who had been diagnosed with these conditions appear to have ldl levels of less than 160 mg/dl.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **LDL<160 mg/dl**  Mean (SD; Min Mdn Max; n) | **LDL>=160 mg/dl**  Mean (SD; Min Mdn Max; n) | **All Patients**  Mean (SD; Min Mdn Max; n) |
| **Age (yrs)**  **Sex**  **Male**  **Female**  **Weight (pounds)**  **Smoking History**  **Chd**  **0**  **1**  **2**  **Chf**  **0**  **1**  **Stroke**  **0**  **1**  **2** | 74.5 (5.4; 65, 73, 99; n=618)  49 (n=303)  51 (n=315)  159.4(30.8;86, 158, 264; n=618)  19.9 (27.6; 0, 7, 240; n=618)  79 (n=488)  8.7 (n=54)  12.3 (n=76)  94 (n=581)  6 (n=37)  87.5 (n=541)  2.9 (n=18)  9.6 (n=59) | 74.8 (5.8; 65, 74, 94; n=117)  56.4 (n=66)  43.6 (n=51)  163.1(30.4; 74, 158, 257; n=117)  18.1 (24.3;0, 3.8, 102; n=116)  78.6 (n=92)  8.6 (n=10)  12.8 (n=15)  96.6 (n=113)  3.4 (n=4)  81.2 (n=95)  5.1 (n=6)  13.7 (n=16) | 74.5 (5.5; 65, 74, 99; n=735)  159.9(30.7; 74, 158, 264; n=735)  19.6 (27.1; 0, 6.5, 240; n=734)  78.9 (n=580)  8.7 (n=64)  12.4 (n=91)  94.4 (n=694)  5.6 (n=41)  86.5 (n=636)  3.3 (n=24)  10.2 (n=75) |

1. In order to compare means of the continuous random variable ldl across patients who survived less than 5 years and those who survived at least 5 years, I perform a t test. In addition, I assume unequal variances between the two groups since the problem doesn’t specify whether to presume equal or unequal variances across the two groups.

The mean ldl is estimated to be 118.70 mg/dl among subjects who survive less than 5 years since the study entry and 127.20 mg/dl among subjects who survive at least 5 years. A comparison of the two groups estimates that the mean ldl is 8.50 mg/dl higher for subjects who survived at least 5 years relative to those who die within 5 years. This observed difference is statistically different from 0 (p=0.0186), with a 95% confidence interval suggesting that the observed difference is what might be typically observed if the true difference between survivors and nonsurvivors was anywhere between 1.44 mg/dl and 15.56 mg/dl, with the survivors averaging higher ldl levels.

Therefore we reject the null hypothesis of no association between survival time and ldl at study entry in favor toward higher mean ldl among subjects surviving longer period of time.

1. To compare the geometric means of ldl across the two groups, I perform a t test on log transformed data, with back transformation of the resulting estimates. I assume unequal variances since the problem doesn’t specify whether to presume equal or unequal variance across the groups.

The geometric mean ldl is estimated to be 112.01 mg/dl among subjects who die within 5 years of study entry and 122.83 mg/dl among subjects who survive at least 5 years. A comparison of the two groups estimates that the geometric mean ldl is 9.65% higher among subjects who survive at least 5 years relative to those who die within 5 years. This observed difference is statistically different from 0 (p=0.0128), with a 95% confidence suggesting that the observed difference is what might be typically observed if the true difference between survivors and nonsurvivors was such that the geometric mean for survivors was anywhere between 2.01% and 17.87% higher than that of nonsurvivors. We thus reject the null hypothesis of no association between survival time and ldl at study entry in favor of a trend toward higher geometric mean ldl among subjects surviving the longer period of time.

1. To compare the probability, I proceed with logistic regression, estimate the log odds from the regression model, exponentiate that to obtain the odds, and then find the probability using prob=odds/(1+odds).

The logistic model of the indicator of death in 5 years on ldl estimates that subjects having an ldl level of 160 mg/dl have an odds of death with 5 years of 0.1569, leading to an estimated probability of death within 5 years of 13.56%.

1. To evaluate the association between serum ldl and all-cause mortality by comparing odds of death within 5 years across groups defined by whether subjects have high serum ldl, I use a logistic regression model of the indicator of early death on the untransformed ldl. This model estimates a common odds ratio for each additive difference in ldl.

When comparing two groups, the odds of dying within 5 years is estimated to be 23.36% lower (odds ratio 0.7664) for each mg/dl difference in ldl level, with the group having the higher level of ldl tending toward a lower odds of death within 5 years. This observed difference is not statistically different from an odds ratio of 1 (p=0.358), with a 95% confidence interval suggesting that the observed odds ratio is what might be typically observed if the true odds of dying within 5 years was anywhere between 56.66% lower and 35.2% higher for each mg/dl ldl level. We therefore fail to reject the null hypothesis of no association between survival time and ldl at study entry in favor of a trend toward higher odds of survival among subjects with higher ldl level.

1. In order to evaluate the association between serum ldl and all-cause mortality over the entire period of observation of the subjects by comparing the instantaneous risk of death across groups, I use a proportional hazards regression model of the censored time to death on the untransformed ldl. This model estimates a common hazard ratio for each additive difference in ldl.

When comparing two groups with different ldl levels, the instantaneous risk of death is estimated to be 25% lower (hazard ratio 0.75) for each mg/dl difference in ldl, with the group having the higher level of ldl tending towards lower instantaneous risk of death. This observed difference is not statistically different from an hazard ratio of 1 (p=0.27), with the 95% confidence interval suggesting that the observed hazard ratio is what might be typically observed if the true instantaneous risk of death was anywhere between 55.03% lower and 25.07% higher for each mg/dl higher in ldl level. We therefore fail to reject the null hypothesis of no association between survival time and ldl at study entry in favor of a trend toward higher odds of survival among subjects with higher ldl level.

1. To address the association between mortality and serum ldl, I would proceed by making the Kaplan Meier survival curves for the two ldl groups (ldl below 160 mg/dl and ldl equal or above 160 mg/dl). I would then use log rank test to test the null hypothesis of equal survival curves for the two groups. I choose to use this nonparameteric test since it’s appropriate for right censored data as it’s the case here. In addition, the censoring is non- informative. The log rank test statistic compares the estimates of hazard functions of the two groups at each observed event time and outputs the p-value which I can use to make a decision on whether to reject or fail to reject with null hypothesis.