**Biostats 518 - Homework #2**

**Question 1**

**1a.**

|  |  |  |
| --- | --- | --- |
|  | **N** | **Mean LDL (SD),**  |
| **Survived through 5 years** | 606 | 127.2 (32.9)  |
| **Died within 5 years** | 119 | 118.7 (36.2) |

606 subjects survived at least 5 years of follow up and 119 did not survive to 5 years of follow-up. The sample mean LDL was higher among those who survived (127.2 mg/dL) compared to those who did not survive to 5 years of follow-up (118.7 mg/dL). The standard deviation was lower for those who survived to at least 5 years (32.9) compared to those who did not survive to 5 years (36.2).

**1b.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **N** | **Mean LDL (SE)** | **95% CI** |
| **Survived through 5 years** | 606 | 127.2 (1.33) | 124.6, 129.8 |
| **Died within 5 years** | 119 | 118.7 (3.31) | 112.1, 125.3 |

Based on the 95% confidence interval, the observed mean LDL of 127.2 mg/dL among subjects who survived at least 5 years would not be unusual if the true population mean were between 124.6 mg/dL and 129.8 mg/dL. Based on the 95% confidence interval, the observed mean LDL of 118.7 mg/dL among subjects did not survive survived to 5 years would not be unusual if the true population mean were between 112.1 mg/dL and 125.3 mg/dL. The standard error of the mean LDL for those who survived to 5 years was smaller than the standard error of the mean LDL for those who did not survive to 5 years (1.33 and 3.31, respectively).

The standard deviation of the mean LDL was 1.1 times higher for those who died compared to those who survived to 5 years. The standard error of the mean LDL was 2.5 times higher for those who died compared to those who survived to 5 years. However, the difference in the standard deviations was 3.3 while the difference in the standard error was 1.98. Because the sample size is smaller for those who died within 5 years, we would expect the standard error to be larger compared to those who survived.

**1c.** The 95% CIs comparing those who survived and those who did not survive to 5 years do overlap. However, no conclusions regarding the statistical significance of the difference in mean LDLs can be reached from this observation because the mean LDL for one group is not is included in the 95% CI of the other group.

**1d.** Assuming equal variances between the 2 groups, it would be appropriate to calculate the combined standard deviation using the average standard deviations of the 2 groups weighted by their sample size. The combined standard deviation using this method is 33.6.

**1e.** Based on 95% confidence interval assuming equal variances, the observed lower mean LDL of 8.50 mg/dL among those who died within 5 years compared to those who survived to 5 years would not be unusual if the true difference in the population mean LDL were between the 95% confidence intervals of 1.9 mg/dL and 15.1 mg/dL lower. The t-test assuming equal variance shows that this difference in mean LDL is statistically significant based on an alpha level of 0.05 (two-sided p-value=0.0115). Thus, we reject the null hypothesis of no difference in the mean LDL between those who died and those who did not die within 5 years of follow-up.

**Question 2**

**2a.** Both models are saturated because in both models, there are only 2 values of the predictor of interest possible (either yes or no to surviving or dying within 5 years). The number of groups equals the number of parameters.

**2b.** Using the intercept from model A where the predictor of interest is dying within 5 years, the estimate of the true mean LDL is 127.198 mg/dL for those who survived to 5 years. This is identical to the estimate for the mean LDL for those who survived found when performing a t-test in question 1 that assumed equal variance.

**2c.** Using regression model A, the 95% confidence interval for the true mean LDL among those who survived to 5 yearsis 124.571 to 129.8679. In problem 1, the 95% confidence interval for the true mean LDL among those who survived to 5 yearsis 124.5282 to 129.8679. These estimates are very similar, though the slight difference between the 2 confidence interval estimates is due to slight differences in the variance estimates. Because the variance is assumed to be equal, even though this is not exactly true, the estimates of the standard error, and therefore the confidence intervals, will differ between the two methods.

**2d.** Using the intercept from model B where the predictor of interest is surviving at least 5 years, the estimate of the true mean LDL is 118.7 mg/dL for those who died within 5 years. This is identical to the estimate for the mean LDL for those who died found when performing a t-test in question 1 that assumed equal variance.

**2e.** Using regression model A, the 95% confidence interval for the true mean LDL among those who survived to 5 yearsis 112.6726 to 124.7224. In problem 1, the 95% confidence interval for the true mean LDL among those who survived to 5 yearsis 112.1338 to 125.2611. These estimates are very similar, though the slight difference between the 2 confidence interval estimates is due to slight differences in the variance estimates. Because the variance is assumed to be equal, even though this is not exactly true, the estimates of the standard error, and therefore the confidence intervals, will differ between the two methods.

**2f.** Assuming equal variances between the two populations, the root MSE would be the appropriate method to estimate the standard deviation. For both models A and B, the root MSE is 33.477. For problem 1, the combined standard deviation is 33.6. These numbers are very similar. The slight difference is due to the assumption of equal variance that does not hold perfectly in these models.

**2g.** Models A and B are inverses of each other. They provide almost exactly the same information, and are the same model, except that in model A, the intercept is the mean LDL for those who survive to 5 years, and in model B the intercept is the mean LDL for those who died within 5 years. The absolute value of the slope is the difference between the two means in both models with the same standard error (with the same absolute values for the 95% confidence intervals).

**2h.** Based on the 95% confidence interval, the observed mean LDL of 127.2 mg/dL among subjects who survived at least 5 years would not be unusual if the true population mean were between 124.5 mg/dL and 129.9 mg/dL. The intercept of model A is the mean LDL among those who survived to 5 years of follow-up.

**2i.** Based on the 95% confidence interval, the observed difference in mean LDL of -8.5 mg/dL between subjects died within 5 years and those who survived at least 5 years would not be unusual if the true population difference in the mean were between -15.1 and -1.9 mg/dL. The slope of model A is the difference between the mean LDL comparing those who died to those who survived to 5 years or follow-up.

**2j.** The true difference in the mean LDL comparing those who survived at least 5 years to those who died within 5 years was 8.5 mg.dL. Based on the 95% confidence interval, this observed mean difference would not be unusual if the true population difference in the mean were between 1.9 mg/dL and 15.1 mg/dL. The p-value testing the hypothesis that the two populations have the same mean LDL is 0.0115. Based on an alpha level of 0.05, we can conclude that this is a statistically significant difference in mean LDL, and that the groups do not have the same LDL. We reject the null hypothesis that the groups do not differ by mean LDL.

**Question 3**

Using a t-test allowing for unequal variance, the mean difference in LDL comparing those who survived to 5 years to those who did not survive to 5 years was 8.5 mg/dL, with a standard error of 3.57 and 95% CIs of 1.44 mg/dL to 15.6 mg/dL. Compared to question 1 that assumed equal variances, the difference in the mean is the same, though the standard error is somewhat smaller at 3.36 (95% CI: 1.9 mg/dL , 15.1 mg/dL; two-sided p-value=0186). Assuming equal variances in question 1 resulted in confidence intervals that were too small (anti-conservative). However, the end result is similar in that in both t-tests, the two-sided p-values were less than 0.05, resulting in rejecting the null hypothesis that the mean LDL was the same in the two groups.

**Question 4**

Using a linear regression model that allows for the possibility of unequal variances (robust), the slope and intercept remains that same as in question 2 though the standard errors are larger and the 95% CIs are wider. However, the interpretation remains the same because the p-value for the slope (difference in the mean LDL levels) is less than 0.05 in both methods. Comparing the results of the regression with robust variance and a t-test that allows for unequal variance, the standard error for the difference in the mean LDL between groups is very similar for both methods (3.57). In both methods, the p-value is less than 0.05, so we can reject the null hypothesis that the mean LDL is the same in the two groups. These p-values are also very similar for these methods than allow for unequal variances (t-test p-value=0.0186; regression p-value: 0.017).

**Question 5.**

**5a.** The mean age is similar between LDL category levels (Table 1). More males than females are in the below 130 mg/dL LDL category, while more females than males are in the categories over 130 mg/dL (Table 1). When stratifying the mean age by both LDL and sex (Table 3), the mean age was similar within and across sex and LDL categories (approximately 74-75 years). The mean LDL appears to be decreasing slightly with increasing age, although this is not completely linear with an increase in LDL in the 80-85 age group (Table 2).

**Table 1**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Below 130 mg/dL (n=393)** | **130 to 159 mg/dL (n=225)** | **160 mg/dL and up (n=107)** | **Total** **(n=725)** |
| **Mean Age in years****(sd; min-max)** | 74.7 (5.25; 65 - 92) | 74.2 (5.62; 67 – 99) | 74.9 (5.8; 65 – 94) | 74.6 (5.44; 65-99) |
| **Percent male** | 55.5% | 43.1% | 42.1% | 49.7% |

**Table 2**

|  |  |  |  |
| --- | --- | --- | --- |
| **Age Category** | **N** | **Mean LDL (sd; min - max)** | **Percent Male** |
| **65-74** | 174 | 130.0 (32.9; 51 – 247) | 48.3% |
| **70-75** | 295 | 123.4 (33.6; 37 – 216) | 49.5% |
| **75-80** | 155 | 124.7 (33.3; 11 – 225) | 52.3% |
| **80-85** | 66 | 129.4 (33.0; 69 – 227) | 43.9% |
| **85+** | 35 | 123.4 (38.6; 57 – 216) | 57.1% |
| **Total** | 725 | 125.8 (33.6; 11 – 247) | 48.3% |

|  |  |  |
| --- | --- | --- |
|  | **Females** | **Males** |
| **LDL category (mg/dL)** | **N** | **Mean Age (sd; min – max)** | **N** | **Mean Age (sd; min – max)** |
| **Below 129** | 175 | 74.3 (5.16; 65 - 89) | 218 | 75.0 (5.31; 66 - 92) |
| **130 to 159** | 128 | 74.4 (5.15; 67 - 91) | 97 | 74.0 (6.21; 67 - 99) |
| **160 and up** | 62 | 74.8 (5.80; 65 - 89) | 45 | 74.9 (5.79; 67 - 94) |
| **Total** | 365 | 74.4 (5.26; 65 - 91) | 360 | 74.7 (5.63; 66 - 99) |

**Table 3**

**5b.** Linear regression was performed using robust standard error estimates to determine if LDL is associated with age in a population of subjects aged 65 to 99 years.

**5c.** This is not a saturated model. Age, the predictor of interest, is a continuous variable, and therefore the number of groups is much greater than the number of parameters.

**5d.** The regression model was found to be LDL=(-0.09)\*Age+132.53

 For subjects 70 years old the estimated mean LDL is 126.21 mg/dL.

**5e.** For subjects 71 years old the estimate mean LDL is 126.12 mg/dL.

The slope is equal to the difference in mean LDL between subjects one year of age apart. Thus, the difference between subjects who are 71 compared to those who are 70 is equal to the slope which is -0.0901, meaning those who are 71 years old have an estimated mean LDL that is 0.0901 mg/dL lower than the mean LDL of those who are 70 years old.

**5f.** For subjects 75 years old the estimate mean LDL is 125.76 mg/dL.

The slope is equal to the difference in mean LDL between subjects one year of age apart. Thus, the difference between subjects who are 75 compared to those who are 70 is equal to 5 times the slope (-0.0901) which is -0.45, meaning those who are 75 years old have an estimated mean LDL that is 0.45 mg/dL lower than the mean LDL of those who are 70 years old.

**5g.** The root MSE was 33.6 in this regression model. This within group standard deviation is 33.6.

**5h.** The intercept in this model is referring to the LDL level of someone at age zero (newborns). In this situation where the subject population is limited to those age over 65 years, this intercept does not have a relevant scientific interpretation. It would be invalid to extrapolate these trends seen in the elderly to children.

**5i.** The slope is equal to the difference in mean LDL between subjects of one age compared to subjects 1 year younger.

**5j.** Using linear regression with robust standard error estimates to detect an association between LDL and age, we found that the difference in mean LDL between subjects of one age compared to subjects 1 year younger was -0.09. Based on the 95% confidence interval, the observed difference in mean LDL of -0.09 mg/dL between subjects one year of age apart would not be unusual if the true difference were between -0.547 mg/dL and 0.367 mg/dL. The p-value for the slope was 0.968, so we therefore fail to reject the null that this difference in LDL by age is zero. There is insufficient evidence to conclude that LDL is associated with age.