HW 2

Biost 518

Scott Emerson

1/21/14

1. A dummy variable for mortality was created. Observations that had a recorded death within 5 years were coded as 1, and observations that had a follow-up time greater than 5 years were coded as 0. The sample size was determined, as well as the sample mean and standard deviation. Along with point estimates for the true mean LDL, the standard error and 95% confidence intervals were calculated for each mortality group. The true difference in mean LDL by survival populations was also estimate, as well as a corresponding standard error and 95% confidence interval. A two-sided, two sample t-test assuming equal variances was also preformed. The test was conducted to assess the null hypothesis that mean LDL is equal between observations that died within 5 years and observations that had survived after 5 years. That is, the null hypothesis that difference in the mean LDL between the two mortality groups is equal to 0 was assessed.

	1. The sample size, sample mean, and sample standard deviation of LDL for each mortality group are presented in the table below. The mean LDL for the group that died within 5 years is lower than the mean LDL of the group that survived past 5 years, however the magnitude of the difference is not that large. The sample standard deviations are similar in both groups. The sample size in the group that survived past 5 years is much larger than that of the subjects that died within 5 years.

|  |  |  |
| --- | --- | --- |
| LDL Summary Statistic | Death ≤ 5 Years | Death > 5 Years |
| Sample Mean (mg/dL) | 118.70 | 127.20 |
| Sample Std. Deviation (mg/dL) | 36.16 | 32.93 |
| Sample Size | 119 | 606 |

* 1. For each survival group, the point estimate and its estimated standard error, and the 95% confidence interval for the true mean LDL are presented in the table below. The point estimates for the true mean are same as the sample means reported above. Again, although the mean LDL for the subjects that died within 5 years is lower, the magnitude of the point estimates is not that different. The standard errors are quite different in magnitude. The standard error is much larger for the subjects who died within 5 years. The difference in the magnitude for the standard error estimates is due to the difference in sample sizes. The large sample size in the survival group is much larger, causing the standard error to be much smaller.

|  |  |  |
| --- | --- | --- |
| LDL Estimate | Death ≤ 5 Years | Death > 5 Years |
| Point Estimate (mg/dL) | 118.70 | 127.20 |
| Standard Error | 3.31 | 1.34 |
| 95% Confidence Interval | (112.13, 125.26) | (124.57, 129.83) |

* 1. ls for the two groups overlap. Based on this observation I would say we the difference in the means is not statistically significant. That is, we cannot say with certainty the difference in mean LDL for the group that died within 5 years and the group that survived past 5 years is not, is not 0.
	2. If we assume equal variances a pooled standard deviation estimate should be used. The pooled standard deviation for these data is 33.60. The formula for the pooled standard deviation is shown below. Sx and Sy are the sample standard deviations for the each group and n1 and n2 are the sample sizes for each group.
	3. The point estimate for the true difference of mean LDL for subjects who died within 5 years compared to subjects who survived past 5 years is 8.50. The estimated standard error for this point estimate is 3.37. The 95% confidence interval for the true difference in mean LDL is (1.91, 15.09). There is a 95% chance this confidence interval. Based on this confidence interval, observed values of mean LDL would not be unusual if the true difference in population means were anywhere between 1.91 and 15.05 mg/dL. Note the confidence interval is positive and does not contain 0. The p-value yielded from the two-sample t-test assuming equal variances is 0.01. Given the data we observe, the probability that the true difference in mean LDL between mortality groups is 1%. Given this p-value, we reject the null hypothesis. Based on all of these analyses there is a statistically significant association between LDL and 5 year all cause mortality. Subjects who survived longer than 5 years have higher serum LDL levels on average. However, this does not agree with the conclusions reached in (c).
1. Using ordinary least squares, a line was fitted to determine the association between LDL and 5-year all cause mortality. Classical regression was used, that is, homoscedasticity was assumed. A dummy variable for mortality was created. Observations that had a recorded death within 5 years were coded as 1, and observations that had a follow-up time greater than 5 years were coded as 0. Regression was used to fit LDL on the dummy variable, to assess the association between LDL and death within 5 years. A second line was fitted, again with LDL as the outcome and 1 minus the dummy variable was the predictor of interest. This line was fitted to determine the association between LDL and survival after 5 years.

	1. Both of these models are saturated. In either model, there are only two parameters in the model (LDL and the mortality dummy variable). In our predictor of interest there are only two groups (those that died within 5 years and those that survived more than 5 years), so the number of parameters is the same as the number of the groups, for both models.
	2. Using the model that fit subject death within 5 years as the predictor of interest, the estimate of true mean LDL among subjects surviving at least 5 years is 127.20 mg/dL. This is the same as estimated in (1).
	3. Using the model that fit subject death within 5 years as the predictor of interest, the confidence interval for true mean LDL among subjects surviving at least 5 years is (124.53 mg/dL, 129.87 mg/dL). This confidence interval is slightly more conservative (larger) than what was computed in (1). The difference in confidence interval is due the standard deviation, and therefore standard error used the confidence interval calculation. In (1) the one sample standard deviation was used. In regression, the pooled standard deviation is used.
	4. Using the model that fit subject survival after 5 years as the predictor of interest, the point estimate of the true mean LDL for subjects who died within 5 years is 118.70 mg/dL. This is the same estimate as found in (1).
	5. Using the model that fit subject survival after 5 years as the predictor of interest, the confidence interval for the true mean LDL for subjects who died within 5 years (112.67 mg/dL, 124.72 mg/dL). The confidence interval from regression is slightly less conservative (smaller) compared to the one found using the t-test. This is because the regression-based confidence interval uses a pooled standard deviation in the confidence interval calculations.
	6. When using regression and assuming equal variances in the two populations the pooled standard deviation is used as the within group standard deviation for both models. The formula for the pooled standard deviation is shown below. This is different than the classic standard deviation formula that is used to calculate to the standard deviation separately for each group as in (1).
	7. The models in A and B are the same. The only difference is a reparameterization of the predictor of interest, which in both cases is a binary variable that indicates survival at 5 years. They are both modeling the relationship between 5 year all cause mortality and LDL.
	8. In model, A the intercept represents the estimated mean LDL level for subjects who survived after 5 years.
	9. The slope represents the difference in mean LDL levels between subjects who died within 5 years and subjects who survived after 5 years.
	10. The point estimate for the true difference in mean LDL levels between the two populations is 8.50 mg/dL. The standard error for the point estimate is 3.36. The 95% confidence interval for the true difference in mean LDL is
	(1.91 mg/dL, 15.09 mg/dL). The p-value to test they hypothesis that the mean LDL levels are the same between the two populations is 0.01. Based on all of these results, there appears to be a relationship between 5 year all cause mortality and LDL level. Subjects who survived more than 5 years had higher LDL levels on average. The conclusions reached from this analysis are the same as reached in (1).

1. A two-sided, two-sample t-test, assuming unequal variances, was preformed to assess the relationship between 5 year all-cause mortality and serum LDL levels. Means, standard errors, and 95% confidence intervals were calculated for each mortality group to estimate the true mean LDL levels by survival group. The true difference in mean LDL level was also estimate and its standard error as well as 95% confidence interval was also constructed to help make inference on the association between survival and LDL levels.

The sample mean, sample standard deviations, and samples sizes are the same as reported in (1a). For the subjects who died within one year, the estimate for the true mean LDL level is 118.70 mg/dL. The standard error for this estimate is 3.31, and the corresponding 95% confidence interval is (112.13 mg/dL, 125.26 mg/dL). These estimates are all the same as found in (1). For the subjects who survived after 5 years, the estimated true mean LDL level is 127.20 mg/dL. The standard error estimate is 1.34 and the corresponding 95% confidence interval is (124.57 mg/dL, 129.83 mg/dL). Again, these estimates are the same as found in (1). The point estimate for the true difference in mean LDL levels between the subjects who died within 5 years and subjects who survived after 5 years is 8.50 mg/dL. The standard error for this estimate is 3.57. The corresponding 95% confidence interval for the true difference in mean LDL is (1.44 mg/dL, 15.56 mg/dL). The estimate for the difference in means is the same as found in (1). However, the standard error for the point estimate is larger than that found in (1). The confidence interval for the true difference in mean LDL between the two populations is also larger than what was found in (1). The p-value yielded from the two-sample t-test is 0.02. Therefore, we reject the null hypothesis and find there is an association between LDL and 5 year all-cause mortality. Based on all the information above, it appears that subjects who died within 5 years have lower LDL levels compared to subjects who survived after 5 years. Although the p-value yielded from this t-test is different that what was found in (1) the conclusion reached is the same.

1. Using ordinary least squares, a line was fitted to determine the association between LDL and 5-year all cause mortality. Robust regression was used, that is, heteroscedasticity was assumed. Two parameterizations of the model were fitted. In both cases the response was LDL. In the first model the predictor of interest was subject death within 5 years. In the second model the predictor of interest was subject survival after 5 years. Using the regression results the true mean LDL levels were estimated for each group. Standard error estimates were also calculated for the point estimates for each group. A 95% confidence interval was also constructed for the true mean LDL level for each mortality group. The true difference in mean LDL levels was also estimated using the regression results. A corresponding standard error and 95% confidence interval were also yielded from the model.

For subjects who died within 5 years the estimate for the true mean LDL is 118.70 mg/dL. The standard error estimate for this point estimate is 3.31 mg/dL. The 95% confidence interval for the true mean LDL for subjects who died within 5 years is (112.21 mg/dL, 125.19 mg/dL). The point estimate for mean LDL and the estimate standard error are the same as in (3). The confidence interval found using robust regression is slightly larger (more conservative) than the one found using the t-test assuming unequal variances in (3). The point estimate for the true mean LDL level for subjects who survived after 5 years is 127.198 mg/dL. The estimated standard error for this point estimate is 1.34. The 95% confidence interval for the true mean LDL level for subjects who survived after 5 years is (124.57 mg/dL, 129.83 mg/dL). The point estimate for mean LDL level for subjects who survived after 5 years and the estimated standard error found using robust regression are the same as found in (3). The confidence interval found using robust regression is the same that was found using the t-test for unequal variances in (3). The estimate for the true difference in mean LDL levels between mortality groups is 8.50 mg/dL. The standard error for the estimated difference is 3.57. The corresponding 95% confidence interval for the true difference in mean LDL levels between populations is (1.50 mg/dL, 15.50 mg/dL). The point estimate for the true difference in mean and standard error estimate found using robust regression are the same as found in (3). However, the confidence interval found using robust regression is slightly smaller (more conservative) than the one found using the t-test for unequal variances in (3). The p-value yielded in robust regression to assess if the difference in mean LDL for each population is 0 is 0.02. Therefore, we reject the null hypothesis that the mean LDL level is the same between subjects who died within 5 years and subjects who survived after 5 years. Further, it appears the subjects who died within 5 years have lower LDL levels compared to subjects who survived after 5 years. The p-value and overall conclusions reached are similar to what was found in (3).

1. Using ordinary least squares, a line was fitted to determine the association between LDL and age, where age is treated as continuous variable. Classical regression was used, that is, homoscedasticity was assumed. Using the regression results the true mean LDL levels were estimate for a 1-year increase in age. Standard error estimates were also calculated for the point estimates for each group. A 95% confidence interval was also constructed for the true mean LDL level for 1-year increase in age. Descriptive statistics were also generated for LDL for the whole sample as well as by categorized age. The descriptive statistics for sex were also generated for the whole sample and by age category to assess for confounding.
	1. The table of descriptive statistics is shown below. The mean LDL levels appear to be the most the same across age groups. The standard deviation of LDL is larger for the oldest age category, however is likely due to the small sample size. The distribution of gender is similar for the lower age ranges, with the proportion of men and women being roughly equal. However there are a higher percentage of men (60%) in the oldest age category (85+).

|  |
| --- |
| **Table I. Summary Statistics By Age Group** |
| Variables | All (N=735) | 65-74 Years (N=417) | 75-84 Years (N=264) | 85+ Years (N=44) |
| Mean  | SD | Min, Max | Mean  | SD | Min, Max | Mean  | SD | Min, Max | Mean  | SD | Min, Max |
| LDL (mg/dL) | 125.80 | 33.60 | 11, 247 | 125.97 | 32.45 | 37, 247 | 125.63 | 34.87 | 11, 227 | 125.25 | 37.23 | 57, 216 |
| Gender |  |  |   |   |  |   |  |  |   |   |  |  |
| Male | 0.498 | - | - | 0.491 | - | - | 0.493 | - | - | 0.600 | - | - |
| Female | 0.502 | - |  - | 0.509 | - | - | 0.507 | - | - | 0.400 | - | - |

* 1. Ordinary Least Squares regression was used to fit a line that models the association between LDL and age. Classical regression was used, and therefore homoscedascity was assumed. Age was treated as a continuous variable.
	2. This model is not saturated. Because age is treated as a continuous variable, each observed age value is treated as a group. There are only two parameters in the mode, but there are many more groups in the model (age ranges from 65-99). A saturated model is a model where the number of groups and the number of parameters are the same, which is not true in this model.
	3. The estimated mean LDL level for a population of 70 year subjects is 126.21 mg/dL.
	4. The estimated mean LDL level for a population of 71 year old subjects is 126.12 mg/dL. This estimate is 0.09 mg/dL lower than the estimate of mean LDL for the population of 70-year old subjects. That difference in these estimates is the same as the slope estimated early.
	5. The estimated mean LDL level for a population of 75 year old subjects is 125.76 mg/dL. The difference between this estimate and the one found for 70 year subjects is -0.45 mg/dL. This is exactly 5 times the estimated slope, which is the difference in years of the two estimates multiplied by the estimated slope.
	6. The root mean squared error for the classical regression model is the pooled standard deviation for each group. That is, it is the pooled standard deviation for all the different ages observed.
	7. The intercept can be interpreted as the mean LDL level for the population of subjects who are 0 years old. This does not have any relevant scientific meaning, as we are interested in the effects of LDL on an older population.
	8. The slope estimates the true difference in mean LDL for a one-year age difference. For example, the slope estimates the difference in mean LDL level between a population of subjects who are 80 years old an a population of subjects who are 81 years old.
	9. The estimate for the true mean difference in LDL levels for populations one year apart in age is -0.09 mg/dL. The estimated standard error for this estimate is 0.229 mg/dL. The 95% confidence interval for the true difference in mean LDL for populations one year apart is (-0.541 mg/dL, 0.360 mg/dL). An estimated difference in mean LDL between populations on year apart that falls into this interval would not be considered unusual. Note that this confidence interval contains 0. The p-value used to test the hypothesis that the true difference in mean LDL for populations on year apart is equal to 0 is 0.694. Therefore, we fail to reject the null that the true difference in mean LDL levels between populations on year apart is 0. That is, we cannot say with certainty that there is a difference in mean LDL levels for different age group populations. This finding is consistent with the confidence interval yielded.
	10. The true difference in mean LDL across groups that differ by 5 years can be estimated by multiplying the slope found in previous model by 5. To construct the 95% confidence interval for the true difference in mean LDL across groups that differ by 5 years I would use the standard error that was estimated in the previous model and the slope from the previous model multiplied by 5 in the standard confidence interval formula.
	11. The estimate correlation between LDL and age is -0.015. The p-value associated with the test of the null hypothesis that the correlation is 0 is 0.694. Therefore, we reject the null hypothesis and cannot conclude that the correlation between age and LDL is nonzero. This is the same conclusion that is reached when assessing the relationship between LDL and age through regression. Further, the p-values yielded in each test are exactly the same.