Biost 515 Homework 2

1. Recall that dichotomization of mortality by whether it occurs before or after 5 years is possible due to the fact that there is no censoring event before 5 years (from homework 1).
2. The sample size, sample mean, and sample standard deviation of LDL values among subjects who survived at least 5 years are 606, 127.198, and 32.929, respectively. The sample size, sample mean, and sample standard deviation of LDL values among subjects who died within 5 years are 119, 118.698, and 36.157, respectively. The sample means are significantly different in magnitude since the difference of sample means between survivors and nonsurvivors is 8.5 mg/dL. The standard deviations (32.929 mg/dL for survivors and 36.157 mg/dL for nonsurvivors, or about 3.2 mg/dL higher for nonsurvivors) are also considerably different since we observe a difference of 3.2 mg/dL.

|  |  |  |  |
| --- | --- | --- | --- |
| **Descriptive Statistics of LDL (based on 5 years)** | | | |
|  | **N** | **mean** | **sd** |
| **suvivors** | 606 | 127.198 | 32.929 |
| **nonsurvivors** | 119 | 118.698 | 36.157 |
| **Total** | 725 | 125.803 | 33.602 |

1. We can use the distribution of sample mean LDL to estimate the distribution of true mean LDL. The mean of the distribution of sample mean is just the sample mean itself, but its standard error is the standard deviation divided by the square root of the sample size.

Among the group of subjects who would survive at least 5 years, the point estimate of the true mean LDL is 127.198 mg/dL with a standard error of 1.338 mg/dL and 95% confidence interval of (124.571, 129.825).

Among the group of subjects who would die within 5 years, the point estimate of the true mean LDL is 118.698 mg/dL with a standard error of 3.315 mg/dL and 95% confidence interval of (112.134, 125.261).

We recall that each level of risk of cardiovascular disease according to serum LDL levels is stratified by intervals of 30 mg/dL from homework 1. Thus, the point estimates are similar in magnitude since their difference (difference in sample means between survivors and nonsurvivors) of 8.5 mg/dL is not very significant. However, standard errors are not similar in magnitude since it is dependent on the sample size of each group (it is the standard deviation divided by the square root of sample size), and the group of survivors is way larger (n=606) than the group of nonsurvivors (n=119). This results in larger standard error (and thus wider 95% CI) for the group of nonsurvivors relative to the group of survivors.

1. The 95% CI for the mean LDL in a population surviving 5 years (124.571, 129.825) overlaps with the 95% CI for the mean LDL in a population dying within 5 years (112.134, 125.261). For us to be able to make the conclusion that the estimated difference in the estimated means is insignificant, we need the 95% CI for one stratum to contain the point estimate of the other stratum. The point estimates for the true mean LDL were 127.198 mg/dL for the survivors and 118.698 mg/dL for the nonsurvivors. It is obvious that both estimates are not contained in the CI for the other stratum, so we cannot draw any conclusions about the statistical significance of the estimated difference in the estimated means.
2. Since we presume that the variances are equal in the two populations, it follows that the standard deviations are equal in the two populations. We can use the pooled standard deviation (square root of the weighted average of the variances) to combine the standard deviation of the two groups. Based on the regression output, we can see that the pooled standard deviation is 33.602 mg/dL.
3. Using the two-sample t-test assuming equal variances between groups, the point estimate of the true difference in means between a population that survives at least 5 years and a population that dies within 5 years is 8.501 mg/dL, with an estimated standard error of 3.357 mg/dL. The 95% CI for the true difference in means is (1.911, 15.090) where the units are measured in mg/dL. Based on the p-value of 0.0115 we can conclude that this observation is statistically significant at the 0.05 level, and with high confidence can reject the null hypothesis that the mean LDL levels are not different by death status at 5 years in favor of a hypothesis that death within 5 years is associated with lower mean levels of LDL.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Two-sample t-test with equal variances** | | | |  |  |  |
| **Group** | **Obs** | **Mean** | **Std. Err.** | **Std. Dev.** | **[95% Conf.** | **Interval]** |
| **Survivors** | 606 | 127.198 | 1.338 | 32.929 | 124.571 | 129.825 |
| **Nonsurvivors** | 119 | 118.698 | 3.315 | 36.157 | 112.134 | 125.261 |
| **combined** | 725 | 125.803 | 1.248 | 33.602 | 123.353 | 128.253 |
| **diff** |  | 8.501 | 3.357 |  | 1.911 | 15.090 |

**p-value=0.0115**

1. Ordinary least squares regression presuming homoscedasticity is equivalent to linear regression assuming equal variances between groups.
2. Model A: We regress LDL on the indicator variable that the subject died within 5 years.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Regression of LDL on 5 yr. death indicator** | | | |  |  |  |
| **ldl** | **Coef.** | **Std. Err.** | **t** | **P>** **|t|** | **[95% Conf.** | **Interval]** |
| **5 yr. death** | -8.501 | 3.357 | -2.530 | 0.012 | -15.090 | -1.911 |
| **intercept** | 127.198 | 1.360 | 93.530 | 0.000 | 124.528 | 129.868 |

The model we fit is saturated since the number of groups (two, group of survivors and group of nonsurvivors dichotomized by 5 years) is equal to the number of parameters (two, the intercept and the slope) in our regression model.

Model B: We regress LDL on the indicator variable that the subject survived at least 5 years.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Regression of LDL on 5 yr. survival indicator** | | | |  |  |  |
| **ldl** | **Coef.** | **Std. Err.** | **t** | **P>** **|t|** | **[95% Conf.** | **Interval]** |
| **5 yr. surv** | 8.501 | 3.357 | 2.530 | 0.012 | 1.911 | 15.090 |
| **intercept** | 118.698 | 3.069 | 38.680 | 0.000 | 112.673 | 124.722 |

The model we fit is saturated since the number of groups (two, group of survivors and group of nonsurvivors dichotomized by 5 years) is equal to the number of parameters (two, the intercept and the slope) in our regression model.

1. Using the intercept parameter from Model A in part (a), we can see that the estimate of the true mean LDL among a population of subjects who survive at least 5 years is 127.198 mg/dL. This estimate is equal to the corresponding estimate found in problem 1.
2. Using the intercept parameter from model A we can see that the confidence interval for the true mean LDL among a population of subjects who survive at least 5 years is (124.528, 129.868) where the unit of measurement is mg/dL. This is slightly different from the corresponding estimate from problem 1, which was (124.571, 129.825). The source of this discrepancy arises from the fact that in regression we use the pooled standard deviation, whereas in t-test we use the group standard deviation which resulted in different estimates of standard error.
3. Using the intercept parameter from Model B in part (a), we can see that the estimate of the true mean LDL among a population of subjects who die within 5 years is 118.698 mg/dL. This estimate is equal to the corresponding estimate found in problem 1.
4. Using the intercept parameter from Model B we can see that the confidence interval for the true mean LDL among a population of subjects who die within 5 years is (112.673, 124.722) where the unit of measurement is mg/dL. This is slightly different from the corresponding estimate from problem 1, which was (112.134, 125.261). The source of this discrepancy arises from the fact that in regression we use the pooled standard deviation, whereas in t-test we use the group standard deviation which resulted in different estimates of standard error.
5. If we presume the variances are equal in the two populations, the regression based estimate of the standard deviation within each group for each model is 33.477 mg/dL. This is obtained from looking at the “Root MSE” of the regression output in Stata (according to the lecture notes, this is how the estimates of within group standard deviation is obtained). Compared to the estimate of standard deviation obtained in problem 1 assuming equal variances (33.602 mg/dL), this estimate is slightly different.
6. Model A and Model B deal with the same model, but we are just parameterizing the slope and the intercept differently. Both models can lead to the same inferences.
7. In regression model A, we regressed LDL on the binary variable indicating that the subject died in 5 years. Thus, the intercept of the regression model indicates the estimate of true mean LDL levels for subjects who survive past year 5, which is 127.198 mg/dL.
8. The slope of -8.501 mg/dL indicates the difference between the estimates of true mean LDL levels between subjects who survive past year 5 and subjects who die before year 5. Thus, we can estimate that subjects who die before year 5 have mean LDL levels of 118.698 mg/dL.
9. The point estimate, the estimated standard error of the point estimate, and the 95% confidence interval for the true difference in means between a population that survives at least 5 years and a population that dies within 5 years using regression parameter estimates from part (a) are 8.501 mg/dL, 3.357 mg/dL, and (1.911, 15.090), respectively (where the 95% CI is also measured in mg/dL). The p-value testing the hypothesis that the two populations have the same mean LDL is 0.012 according to the regression output (assuming equal variances). This p-value is significant at the 0.05 level, so with high confidence we can reject the null hypothesis that the survivors and nonsurvivors at 5 years have nondifferent mean LDL levels in favor of the nonsurvivors having lower mean LDL levels.
10. We use a standard t-test with the assumption of unequal variances. The results of this analysis produces the same point estimates, standard errors, and 95% confidence intervals within each group as we found in problem 1. However, although the point estimate of the difference between true mean LDL of survivors and nonsurvivors is the same as that obtained in problem 1, we get a slightly different standard error for the difference (and thus slightly different 95% confidence interval) since we assume unequal variances between the two groups. This is because the t-test assuming equal variances uses pooled standard deviation, whereas we treat standard deviation differently when assuming unequal variances.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Two-sample t-test with unequal variances** | | | | |  |  |
| **Group** | **Obs** | **Mean** | **Std. Err.** | **Std. Dev.** | **[95% Conf.** | **Interval]** |
| **Survivors** | 606 | 127.198 | 1.338 | 32.929 | 124.571 | 129.825 |
| **Nonsurvivors** | 119 | 118.698 | 3.315 | 36.157 | 112.134 | 125.261 |
| **combined** | 725 | 125.803 | 1.248 | 33.602 | 123.353 | 128.253 |
| **diff** |  | 8.501 | 3.574 |  | 1.441 | 15.560 |
| **p-value = 0.0186** |  |  | |  |  |  |

1. We use linear regression with the assumption of unequal variance. The results we obtain from this analysis do not differ significantly from our results in problem 3. The point estimates are the same, but the standard error and 95% CI differs slightly because using robust standard error is an approximation to the t-test using the Huber-White Sandwich estimator, not exactly a t-test.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Linear regression assuming unequal variance** | | | | |  |  |
|  |  | **Robust** |  |  |  |  |
| **ldl** | **Coef.** | **Std. Err.** | **t** | **P>|t|** | **[95% Conf.** | **Interval]** |
| **5 yr. death** | -8.501 | 3.566 | -2.380 | 0.017 | -15.501 | -1.500 |
| **constant** | 127.198 | 1.338 | 95.040 | 0.000 | 124.570 | 129.826 |

1. Regression analysis evaluating association between LDL and age:
2. To assess the association between LDL and age, we can use a scatterplot for LDL plotted versus age. There does not seem to be much association between LDL and age since the fitted lines by each gender are fairly constant, and there is a lack of linear relationship between the two variables based on the scatterplot. Based on the fitted lines, we can infer that there is effect modification since it is apparent that being male is associated with lower mean levels of LDL. To assess confounding, we would need to know whether sex is causally associated with LDL levels.

**Scatterplot of LDL levels versus age**



1. We perform a linear regression of LDL on the continuous variable age to explore the mean LDL values as they differ by age. We assume unequal variance in our predictor of interest as there is no reason to assume equal variance, i.e. we use robust standard errors. The intercept of our regression model tells us the expected mean LDL value for a subject who is 0 years old, and the slope tells us the expected rate of change of LDL for a 1 year difference in age.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Linear regression assuming unequal variance** | | | | |  |  |
|  |  | **Robust** |  |  |  |  |
| **ldl** | **Coef.** | **Std. Err.** | **t** | **P>t** | **[95% Conf.** | **Interval]** |
| **age** | -0.090 | 0.233 | -0.390 | 0.698 | -0.547 | 0.367 |
| **constant** | 132.528 | 17.343 | 7.640 | 0.000 | 98.479 | 166.577 |

1. This is not a saturated model. The definition of saturated model is that the number of groups (predictors of interest) should be equal to the number of parameters. However, it is hard to define predictors of interests as groups since it is a continuous variable (although we have two parameters in our regression model which are the slope and the intercept). Thus, it violates the condition that the predictor of interest should be binary in a saturated model. Our model should completely explain our data in a saturated model, but here we are borrowing information to answer questions about the age groups that we don’t have information about.
2. Based on our regression model, the estimated mean LDL level among a population of 70 year old subjects is 126.228 mg/dL.
3. Based on our regression model, the estimated mean LDL level among a population of 71 year old subjects is 126.138 mg/dL. The slope is the estimated change in LDL level per 1 year change in age, so the difference between the answer to this problem and the answer to part d is exactly the slope of our regression model (the difference between estimated LDL levels for 71 year olds and 70 year olds).
4. Based on our regression model, the estimated mean LDL level among a population of 75 year old subjects is 125.778 mg/dL. The difference between the answer to this problem and the answer to part d is exactly five times the slope of our regression model, since it tells us how much estimated mean LDL levels will change when age is changed by 5 years.
5. The root mean squared error is the sample standard deviation of the residuals. If we assume constant variance in age groups, the root mean squared error gives us the estimates of within group standard deviation. However, since we are assuming unequal variance in age groups, the root mean squared error is based on average within-group variance.
6. The intercept, as explained in part b, is the expected mean LDL level for a subject who is 0 years old (technically someone who is just born, estimated here to be 132.528 mg/dL). It does not have a very relevant scientific interpretation because it would be rare for us to be interested in the LDL levels of babies who are just born.
7. The slope is the expected rate of change of mean LDL for subjects differing in age by 1 year (estimated here to be -0.09 mg/dL). Thus, we can interpret that on average, 1 year difference in age is associated with 0.09 mg/dL lower mean LDL in our sample.
8. Based on our regression model, we estimate that when comparing two age groups, the mean LDL level differs on average by 0.09 mg/dL per year difference in age (where higher age is associated with lower LDL level). This inference is based on the slope of our regression model. We also see that the intercept of our regression model is 132.528 mg/dL, but we do not make any inferences on it since it is unfeasible to examine LDL levels of someone who is just born (which is what the intercept means). Based on our p-value of 0.698, we conclude that our test is not significant at the 0.05 and we cannot reject the null hypothesis that there is no difference in mean LDL levels by age. From the 95% CI, we observe that these results would be typical of situations where the true average difference in mean LDL levels were between 0.547 mg/dL lower and 0.367 mg/dL higher per year difference in age. We would push our analysis in favor of a hypothesis preferring the lack of difference between LDL levels and age.
9. The point estimate of the difference in mean LDL across groups that differ by 5 years in age is -0.45 mg/dL. This is obtained by multiplying the slope of our regression model by 5. The 95% CI for the difference in mean LDL across groups that differ by 5 years in age is (-2.735, 1.835). This is obtained by multiplying the confidence interval from the regression output by 5.
10. Based on Stata, the estimated correlation between LDL and age is -0.0146, a result that is not very significant since it is very close to 0. A correlation of 0 indicates the lack of a linear relationship between LDL and age.