BIOST 515/518

Winter 2014

Homework 2

ID 2133

1. (a) Among subjects who survived at least 5 years, the sample size is 606, the sample mean and sample standard deviation of LDL values are 127.20mg/dL and 32.93mg/dL respectively. Among subjects who died within 5 years, the sample size is 119, the sample mean and sample standard deviation of LDL values are 118.70mg/dL and 36.16mg/dL respectively. The sample means are similar in magnitude (those who survived within 5 years have a higher mean) and the sample standard deviations are similar (those who died within 5 years have a higher standard deviation).

(b) In a population of similar subjects who would survive at least 5 years, the point estimate of the true mean is 127.20mg/dL, estimated standard error of the point estimate is 1.34mg/dL, and the 95% confidence interval is (124.57mg/dL, 129.83mg/dL). In a population of similar subjects who would die in 5 years, the point estimate of the true mean is 118.70mg/dL, estimated standard error of the point estimate is 3.31mg/dL, and the 95% confidence interval is (112.13mg/dL, 125.26mg/dL).

The point estimates are similar in magnitude (same as in question 1(a)). The standard error of the point estimate of the true mean for subjects who would die in 5 years is 3 times larger than that of subjects who survive within 5 years. The difference in the magnitude of the standard errors is due to the similar sample standard deviations and different sample sizes. Since standard error is

SE = SD/√N

A smaller sample size of subjects who died within 5 years is reflected in a larger standard error of sample mean.

(c) The (95%) confidence intervals for the mean LDL in a population surviving 5 years and in a population dying within 5 years do overlap. We cannot draw any conclusions since the point estimates of each group are not contained within the confidence intervals of the other group.

(d) If we presume that variances are equal in two populations but want to allow for the possibility that the means might be different, the best estimate for the standard deviation of LDL measurements is to calculate it from the pooled variance formula. With that we get 33.5mg/dL as our standard deviation for each of the two groups.

(e) Define the following:

μ0: Mean serum LDL of population that survives in 5 years

μ1: Mean serum LDL of population that dies within 5 years

D = μ0 - μ1

Then the point estimate for D is 8.50mg/dL, the estimated standard error of the point estimate is 3.36mg/dL, and the 95% confidence interval is (1.91mg/dL, 15.09mg/dL). Thus, based on a 95% confidence interval, the observed difference in mean serum LDL (with the population surviving in 5 years having 8.5mg/dL higher) would not be unusual if the true difference in population means were anywhere between 1.91mg/dL and 15.09mg/dL.

Using a two-sample, two-sided ttest of equal variances, define the hypotheses as follows:

H0: Distributions of both populations are the same

HA: Distributions of both populations are different

Then the two-sided p-value is 0.0115. Using a t test that presumes equal variances,

this observation is statistically significant at a 0.05 level of significance, thus we can conclude with high confidence that the distribution of serum LDL differs between those who do or do not have higher risk of death over a 5-year period. The reason we compare distributions instead of just means here is that any differences in means could be an artifact of differences in variances, since our ttest with equal variances does not allow for that possibility.

2. (a) Model A: Y = β0 + β1X where Y is the estimated mean serum LDL levels, β0 is the estimate mean serum LDL among a population of subjects who survive at least 5 years, X is an indicator variable whether the subject died within 5 years. β1 is the change in estimated mean serum LDL levels between two populations. This is a saturated model. The number of groups here is 2 and the number of parameters here is two. Since number of groups = number of parameters, it is a saturated model.

Model B: Y = β0 + β1X where Y is the estimated mean serum LDL levels, β0 is the estimate mean serum LDL among a population of subjects who died within least 5 years, X is an indicator variable whether the subject survived within 5 years. β1 is the change in estimated mean serum LDL levels between two populations. This is a saturated model. The number of groups here is 2 and the number of parameters here is two. Since number of groups = number of parameters, it is a saturated model.

(b) Using Model A, the estimate of the true mean LDL among a population of subjects who survive at least 5 years is 127.20mg/dL (from the intercept term). This is the same as the corresponding estimate from problem 1.

(c) Using Model A, a confidence interval for the true mean LDL among a population of subjects who survive at least 5 years is (124.53mg/dL, 129.87mg/dL). The confidence intervals here are slightly wider than the ones in problem 1**.** This is because classical linear regression uses Root Mean Squared Error (RMSE), which is pooled standard variance, a different method than in problem 1. Also, the t-statistics (critical value) uses different degrees of freedom.

(d) Using Model B, the estimate of the true mean LDL among a population of subjects who die within least 5 years is 118.70mg/dL (from the intercept term). This is the same as the corresponding estimate from problem 1.

(e) Using Model B, a confidence interval for the true mean LDL among a population of subjects who survive at least 5 years is (112.67mg/dL 124.72mg/dL). The confidence intervals here are slightly narrower than the ones in problem 1**.** This is because classical linear regression uses (Root Mean Squared Error) RMSE, which is pooled standard variance, a different method than in problem 1. Also, the t-statistics (critical value) uses different degrees of freedom.

(f) If we presume the variances are equal in the two populations, the regression-based estimates of the standard deviation within each group are the same, 33.48mg/dL which is the pooled variance, Root Mean Squared Error (RMSE). This is the same as the corresponding estimate in problem 1.

(g) Model A is a reparametrization of Model B. The absolute values of the slopes are the same between two models. The intercept term in Model A is the intercept plus the slope in Model B. In other words, define:

YA: estimated mean serum LDL levels for model A

YB: estimated mean serum LDL levels for model B.

bA: intercept in Model A

bB: intercept in Model B

Among subjects who survived in 5 years, YA = YB + bB = YB – bA

Among subjects who died within 5 years, YB = YA + bA = YA – bB

Thus, the choice of the models is arbitrary.

(h) An interpretation of the intercept from the regression model A is the mean serum LDL levels given survival in 5 years, i.e. E[Serum LDL| survive in 5 years] = 127.2mg/dL.

(i) An interpretation of the slope from the regression model A is the difference in mean serum LDL levels among subjects who died in 5 years compared to subjects who survived in 5 years, i.e. E[Serum LDL| died in 5 years] - E[Serum LDL| survive in 5 years] = -8.50mg/dL.

(j) Using the regression parameter estimates, for the true difference in means between a population that survives at least 5 years and a population that dies within 5 years, the point estimate is 8.50mg/dL, the estimated standard error of the point estimate is 3.36mg/dL, and the 95% confidence interval is (1.91mg/dL, 15.09mg/dL). Thus, based on a 95% confidence interval, the observed difference in mean serum LDL (with the population surviving in 5 years having 8.5mg/dL higher) would not be unusual if the true difference in population means were anywhere between 1.91mg/dL and 15.09mg/dL. The p-value testing the hypothesis that the two populations have the same mean LDL is 0.012. Thus at 0.05 level, we can conclude with high confidence that that the distribution of serum LDL differs between those who do or do not have higher risk of death over a 5-year period. All the values and conclusions are the same as in problem 1.

3. **Methods**: A two-sample, two-sided t-test with unequal variance was performed between mean serum LDL levels and 5 year all-cause mortality rates. 95% confidence intervals for the difference in population means were based on that same handling of variances, as done in STATA.

**Inference**: Of the 606 subjects who survived in 5 years, mean serum LDL was 127.20mg/dL. Of the 119 subjects who died in 5 years, mean serum LDL was 118.70mg/dL. The sample deviations are 32.93mg/dL and 36.16mg/dL respectively for the survived and died groups. The standard errors of the estimates of the population means are 1.34mg/dL and 3.31mg/dL respectively for the survived and died groups. **These are exactly the same as in problem 1.**

Define the following:

μ0: Mean serum LDL of population that survives in 5 years

μ1: Mean serum LDL of population that dies within 5 years

D = μ0 - μ1

Then define the hypotheses as follows:

H0: D = 0

HA: D ≠ 0

The point estimate for D is 8.50mg/dL, the estimated standard error of the point estimate is 3.574mg/dL, higher than that in question 1. Based on a 95% confidence interval, the observed difference in mean serum LDL (with the population surviving in 5 years having 8.50mg/dL higher) would not be unusual if the true difference in population means were anywhere between 1.44mg/dL and 15.56mg/dL The confidence interval here is **wider than that in question 1** (1.91mg/dL 15.09mg/

dL). The two-sided p-value here is 0.0186, **higher than that in question 1**. Thus at 0.05 level of significance, we can conclude with high confidence that the distribution of serum LDL differs between those who do or do not have higher risk of death over a 5-year period.

When there is a difference in variance, the t-test that presumes equal variance is ant-conservative, giving p-values that are too low and CI that are too narrow (since more likely to reject the null). When we allow for unequal variances, we get analyses with higher p-values and wider confidence intervals.

4. **Methods**: Linear regression allowing for the possibility of unequal variances was run on regressing serum LDL on an indicator variable defining vital status at 5 years (indicator = 1 if the subject died in 5 years) with robust standard errors. 95% confidence intervals for the difference in population means were based on that same handling of variances, as done in STATA.

**Inference**: The difference in estimated population mean serum LDL levels is 8.50mg/dL, with subjects who survived in 5 years having a higher mean serum LDL. **These estimates are identical to those in question 3.** The estimated standard error of the estimated population mean is 3.566mg/dL. This is **slightly lower than the estimate in question 3**.

Based on a 95% confidence interval, the observed difference in mean serum LDL (with the population surviving in 5 years having 8.50mg/dL higher) would not be unusual if the true difference in population means were anywhere between 1.50mg/dL and 15.50mg/dL. The confidence interval here is **narrower than that in question 3** (Q3: 1.44mg/dL and 15.56mg/dL).

The two-sided p-value here is 0.017, **smaller than that in question 3** (Q3: 0.0186). Thus at 0.05 level of significance, we can conclude with high confidence that the distribution of serum LDL differs between those who do or do not have higher risk of death over a 5-year period. The narrower and smaller estimates have to do with the different calculations of the standard error and also how the critical values are calculated.

5. (a) Descriptive statistics appropriate to the question of an association between LDL and age that would help evaluate whether any such association might be confounded or modified by sex:

Methods: Age was broken into different 5-year categories as below. The sample sizes, means, standard deviations, and ranges of serum LDL levels for each age-sex combination were reported. Also the proportions of sex at each age group were reported.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Male** | | | **Female** | | | **Total** | |
| Age range (N) | N (%) | **Serum LDL** Mean (SD) | Min – Max | N (%) | **Serum LDL** Mean (SD) | Min - Max | **Serum LDL**  Mean (SD) | Min - Max |
| 65 – 69 (114) | 54 (47.4) | 128.5 (30.84) | 68 – 206 | 60 (42.6) | 126.98 (33.98) | 51 – 217 | 127.70 (32.40) | 51 – 217 |
| 70 – 74 (303) | 150 (49.5) | 119.63 (31.34) | 37 – 188 | 153 (50.5) | 130.90 (33.76) | 46 – 247 | 125.32 (32.50) | 37 – 247 |
| 75 – 79 (184) | 92 (50.0) | 120.18 (32.11) | 39 – 218 | 92 (50.0) | 133.52 (37.53) | 11 – 225 | 126.85 (35.46) | 11 - 225 |
| 80– 84 (80) | 38 (47.5) | 114.63 (35.25) | 52 – 227 | 42 (42.5) | 130.21 (30.36) | 81 – 201 | 122.81 (33.49) | 52 – 227 |
| 85 – 89 (34) | 18 (52.9) | 118.78 (35.20) | 72 – 216 | 16 (47.1) | 131.94 (43.22) | 68 – 216 | 124.97 (39.13) | 68 – 216 |
| 90 – 99 (10) | 8 (80.0) | 122.38 (34.76) | 57 – 175 | 2 (20.0) | 141.50 (0.707) | 141 – 142 | 126.20 (31.70) | 57 – 175 |
| Total (725) | 360 (49.7) | 120.56 (32.15) | 37 - 227 | 365 (50.3) | 130.94 (34.25) | 11 - 247 | 125.80 (33.60) | 11 – 247 |

From the above table, the sample sizes are roughly equal among all age categories between sexes. The mean serum LDL levels tend to be higher among females compared to males for every age group. There is also a difference in distribution of sex among age groups, indicating that sex may be associated with age.

If there is a trend between age and LDL levels for each sex, but the trends differ by a constant, then sex could be an effect modifier. If sex is causally associated with age and serum LDL (based on scientific literature), then sex could be a confounder. The above descriptive statistics gives us support for either of these claims, but not any conclusions about them.

(b) To address the question of an association between LDL and age, a linear regression with robust standard errors (allowing for heteroscedasticity) was run in STATA. A new variable (stdage) was generated in STATA to reparametrize age to the minimum value (65), i.e. stdage = age – 65. Then, a linear regression model was run to regress LDL on stdage.

(c) This is a not saturated model. The number of groups here theoretically infinitely many values since age is a continuous variable, and the number of parameters here is two. Since number of groups ≠ number of parameters, it is not a saturated model.

(d) The estimated mean LDL level among a population of 70 year old subjects, E[Y|X=70], is 126.21mg/dL (126.67-5\*0.09).

(e) The estimated mean LDL level among a population of 71-year old subjects. E[Y|X=71] is 126.12mg/dL (126.67-6\*0.09), is . The slope, β1 = E[Y|X=71] - E[Y|X=70] (i.e. slope is answer in e minus answer in d).

(f) The estimated mean LDL level among a population of 75-year old subjects. E[Y|X=75], is 125.76mg/dL (126.67-10\*0.09),. The slope, β1 = (E[Y|X=75] - E[Y|X=70])/5 (i.e. slope is 5 times the answer in f minus answer in d).

(g) The interpretation of the “root mean squared error” (RMSE) is that the average standard deviation of serum LDL levels within each age group is 33.62mg/dL. Since RMSE assumes homoscedasticity among the groups, it plays no role in calculating the standard errors or giving confidence intervals.

(h) The interpretation of the intercept is the estimated mean serum LDL levels for a person of age 65. There is a possible scientific interpretation here: the mean serum LDL level of a 65-year-old person is 126.67mg/dL.

(i) The interpretation of the slope is the decrease in estimated mean serum LDL levels with the increase of 1 year in age. The estimated mean serum LDL levels decreases by 0.09mg/dL with an increase of 1 year in age.

(j) Methods: Based on a simple linear regression allowing for unequal variances, define the hypotheses as follows:

H0: There is no linear trend between age and serum LDL

HA: There is a linear trend between age and serum LDL

Inference: The estimated trend in mean LDL levels by age is an average difference of -0.0902mg/dL per one-year difference in age (LDL lower in older). The 95% confidence interval for the trend is (-0.547mg/dL, 0.367mg/dL). Thus the observed trend would not be unusual if the true one year difference in population means were anywhere between 0.547mg/dL lower and 0.367mg/dL higher in the older age group. The p-value for the F-test is 0.6984. Thus, we cannot reject the null hypothesis that there is no linear trend between age and serum LDL.

(k) An estimate for difference in mean LDL across groups that differ by 5 years of age is -0.45mg/dL and a 95% confidence interval is (-2.73mg/dL, 1.83mg/dL), using the formula that CI5-year = 5\*CI1-year

(l) Define the hypotheses:

H0: Correlation, ρ, = 0

HA: Correlation, ρ, ≠ 0

The point estimate of the correlation is -0.0146, and the p-value is 0.6944. Thus at the 0.05 level, we cannot reject the null hypothesis that ρ = 0. This conclusion is the same as the regression-based conclusion about an association between LDL and age. Thus the test for correlation is the same as the test for association in simple linear regression.