**Applied Biostatistics II**

Emerson, Winter 2014

**Homework #2**

January 13, 2014

All questions relate to associations among death from any cause, serum low density lipoprotein (LDL) levels, age, and sex in a population of generally healthy elderly subjects in four U.S. communities. This homework uses the subset of information that was collected to examine MRI changes in the brain. The data can be found on the class web page (follow the link to Datasets) in the file labeled mri.txt. Documentation is in the file mri.pdf. See homework #1 for additional information.

1. Perform statistical analyses evaluating an association between serum LDL and 5 year all-cause mortality by comparing mean LDL values across groups defined by vital status at 5 years using a t test that presumes equal variances across groups. Depending upon the software you use, you may also need to generate descriptive statistics for the distribution of LDL within each group defined by 5 year mortality status. As this problem is directed toward illustrating correspondences between the t test and linear regression, you do not need to provide full statistical inference for this problem. Instead, just answer the following questions.
	1. What are the sample size, sample mean and sample standard deviation of LDL values among subjects who survived at least 5 years? What are the sample size, sample mean and sample standard deviation of LDL values among subjects who died within 5 years? Are the sample means similar in magnitude? Are the sample standard deviations similar?

**Answer**: For the 606 study participants surviving 5 years or more, mean serum LDL is 127.2 mg/dL with standard deviation 33.0 mg/dL. Among the 119 participants who died within 5 years of study enrollment, mean serum LDL is 118.7 mg/dL with standard deviation 36.2 mg/dL. Though the group surviving at least 5 years has a mean LDL that is 8.5 mg/dL higher, based on the Mayo Clinic recommendations, the mean values are both in the “Near ideal” range of serum LDL levels (100-129 mg/dL). The group standard deviations are similar in magnitude, though subjects who died within 5 years have a slightly higher standard deviation (about 10% higher).

* 1. What are the point estimate, the estimated standard error of that point estimate, and the 95% confidence interval for the true mean LDL in a population of similar subjects who would survive at least 5 years? What are the corresponding estimates and CI for the true mean LDL in a population of similar subjects who would die within 5 years? Are the point estimates similar in magnitude? Are the standard errors similar in magnitude? Explain any differences in your answer about the estimates and estimated SEs compared to your answer about the sample means and sample standard deviations.

**Answer**: For a population of similar individuals that would survive at least 5 years, a point estimate of true mean serum LDL is 127.2 mg/dL with estimated standard error of 1.3 mg/dL. With 95% confidence, these estimates are consistent with a true mean serum LDL of between 124.6 and 129.8 mg/dL.

In a population of similar subjects that would die within 5 years of enrollment, a point estimate of the true mean serum LDL is 118.7 mg/dL with estimated standard error of 3.3 mg/dL. With 95% confidence, these estimates are consistent with a true mean serum LDL of between 112.1 and 125.3 mg/dL.

These point estimates are again both in the in the Mayo Clinic “Near ideal” range of serum LDL levels (100-129 mg/dL), though subjects surviving at least 5 years have an estimated mean LDL that is 8.5 mg/dL higher than those subjects who would die within 5 years. The estimated standard error for subjects who would survive 5 or more years is over twice the standard error for subjects that would die within 5 years of enrollment. For both of these comparisons, the subjects surviving beyond 5 years again have the higher estimates.

* 1. Does the CI for the mean LDL in a population surviving 5 years overlap with the CI for mean LDL in a population dying with 5 years? What conclusions can you reach from this observation about the statistical significance of an estimated difference in the estimated means at a 0.05 level of significance?

**Answer**: The 95% confidence interval for the mean LDL in a population of subjects surviving at least 5 years overlaps with the 95% confidence interval for subjects who would die within 5 years of enrollment. Based on this overlap, we cannot say that there is statistically significant difference in the estimate means at the 0.05 significance level.

* 1. If we presume that the variances are equal in the two populations, but we want to allow for the possibility that the means might be different, what is the best estimate for the standard deviation of LDL measurements in each group? (That is, how should we combine the two estimated sample standard deviations?)

**Answer**: If we assume that the variances are equal in the two populations, the best estimate for the standard deviation of the LDL measurements in each group is the square root of

[(dfsurvived 5 yrs) SD2survived 5 yrs + (dfdie within 5 yrs) SD2die within 5 yrs] / ( dfsurvived 5 yrs + dfdie within 5 yrs)

= [ ( 118 \* 36.22 + 605 \* 32.92 ) / 723 ]1/2 = 33.5 (approx.)

* 1. What are the point estimate, the estimated standard error of the point estimate, the 95% confidence interval for the true difference in means between a population that survives at least 5 years and a population that dies with 5 years? What is the P value testing the hypothesis that the two populations have the same mean LDL? What conclusions do you reach about a statistically significant association between serum LDL and 5 year all cause mortality?

**Answer**: The point estimate for the true difference in mean serum LDL level between a population that survives at least 5 years and a population that dies within 5 years is 8.5 mg/dL, with estimated standard error of 3.4 mg/dL. The 95% confidence interval for this true difference in mean LDL values is 1.9 to 15.1 mg/dL. For the t-test assuming equal variances in the populations with the hypothesis that the two populations have the same mean LDL, the p-value is 0.0115. Hence we have statistically significant evidence of an association between serum LDL and 5-year all-cause mortality in the population represented by the subjects in our sample.

1. Perform statistical analyses evaluating an association between serum LDL and 5 year all-cause mortality by comparing mean LDL values across groups defined by vital status at 5 years using ordinary least squares regression that presumes homoscedasticity. As this problem is directed toward illustrating correspondences between the t test and linear regression, you do not need to provide full statistical inference for this problem. Instead, just answer the following questions.
	1. Fit two separate regression analyses. In both cases, use serum LDL as the response variable. Then, in model A, use as your predictor an indicator that the subject died within 5 years. In model B, use as your predictor an indicator that the subject survived at least 5 years. For each of these models, tell whether the model you fit is saturated? Explain your answer.

**Answer**: In model A and B there are two groups - one for each value of the indicator. The regression model has two parameters, a slope and an intercept. Since the number of groups equals the number of parameters, both model A and B are saturated.

* 1. Using the regression parameter estimates from one of your models (tell which one you use), what is the estimate of the true mean LDL among a population of subjects who survive at least 5 years? How does this compare to the corresponding estimate from problem 1?

**Answer**: The constant coefficient (i.e. intercept) of model A gives an estimate of true mean serum LDL of 127.2 mg/dL among a population of subjects which survive at least 5 years. This is identical to the estimate of the same quantity in problem 1.

* 1. Using the regression parameter estimates from one of your models (tell which one you use), what is a confidence interval for the true mean LDL among a population of subjects who survive at least 5 years? How does this compare to the corresponding estimate from problem 1? Explain the source of any differences.

**Answer**: Based on the constant coefficient’s confidence interval from model A, a 95% confidence interval for the true mean serum LDL in a population of individuals surviving at least 5 years is 124.5 to 129.9 mg/dL. This confidence interval is wider than the 124.6 to 129.8 mg/dL estimated in problem 1 because the standard error is now being calculated from the pooled sample standard deviation discussed in problem 1 (i.e. root mean squared error) of 33.5 mg/dL instead of the sample standard deviation of the individuals surviving at least 5 years only, which was smaller at 32.9 mg/dL.

* 1. Using the regression parameter estimates from one of your models (tell which one you use), what is the estimate of the true mean LDL among a population of subjects who die within 5 years? How does this compare to the corresponding estimate from problem 1?

**Answer**: The constant coefficient (i.e. intercept) of model B gives an estimate of true mean serum LDL of 118.7 mg/dL among a population of individuals surviving less than 5 years. This estimate is identical to the point estimate from problem 1.

* 1. Using the regression parameter estimates from one of your models (tell which one you use), what is a confidence interval for the true mean LDL among a population of subjects who die within 5 years? How does this compare to the corresponding estimate from problem 1? Explain the source of any differences.

**Answer**: Based on the constant coefficient’s confidence interval from model B, a 95% confidence interval for the true mean serum LDL in a population of individuals surviving at least 5 years is 112.7 to 124.7 mg/dL. This confidence interval is narrower than the 112.1 to 125.3 mg/dL estimated in problem 1 because the standard error is now being calculated from the pooled sample standard deviation discussed in problem 1 (i.e. root mean squared error) of 33.5 mg/dL instead of the sample standard deviation of the individuals surviving less than 5 years only, which larger at 36.2 mg/dL.

* 1. If we presume the variances are equal in the two populations, what is the regression based estimate of the standard deviation within each group for each model? How does this compare to the corresponding estimate from problem 1?

**Answer**: Assuming equal variances in the two populations, the regression-based estimate of standard deviation within each group is 33.5 mg/dL in both models A and B. This is identical to the combined standard deviation estimate from problem 1.

* 1. How do models A and B relate to each other?

**Answer**: Model A evaluated at 0 is equal to model B evaluated at 1, and model B evaluated at 0 is equal to model A evaluated at 1. This occurs because the indicator predictor used for model A is equal to one minus the indicator predictor for model B.

* 1. Provide an interpretation of the intercept from the regression model A.

**Answer**: The intercept of regression model A estimates the true mean serum LDL level in a population of subjects that would live at least 5 years.

* 1. Provide an interpretation of the slope from the regression model A.

**Answer**: The slope of regression model A is the estimated difference in the population mean serum LDL value between a population of subjects that would live at least 5 years and a population of subjects that would die within 5 years.

* 1. Using the regression parameter estimates, what are the point estimate, the estimated standard error of the point estimate, 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? What is the P value testing the hypothesis that the two populations have the same mean LDL? What conclusions do you reach about a statistically significant association between serum LDL and 5 year all cause mortality? How does this compare to the corresponding inference from problem 1?

**Answer**: Based on the slope parameter estimate from either model (the models differ only in the direction of the estimated difference), the point estimate for the true difference in means between a population that survives at least 5 years and a population that dies within 5 years is 8.5 mg/dL with a standard error of 3.4. The 95% confidence interval for the true difference in means is 1.9 to 15.1 mg/dL. For the hypothesis that the two populations have the same mean LDL, the p-value is 0.012. Hence the observed difference of 8.5 mg/dL is statistically significant, so we have evidence of an association between serum LDL and 5-year all-cause mortality. This conclusion, the estimates, and p-value are the same as in the inference from problem 1.

1. Perform statistical analyses evaluating an association between serum LDL and 5 year all-cause mortality by comparing mean LDL values across groups defined by vital status at 5 years using a t test that allows for the possibility of unequal variances across groups. How do the results of this analysis differ from those in problem 1? (Again, we do not need a formal report of the inference.)

**Answer**: When performing the t-test that allows the possibility of unequal variances across groups, the sample size, mean, and standard deviation for each group all remain the same. Likewise the estimated standard error and 95% confidence interval for each group are unchanged. Thus, the point estimate for the difference in population mean serum LDL between the groups defined by vital status at 5 years is still 8.5 mg/dL. However, the estimated standard error of the difference in means has increased to 3.6 mg/dL by allowing for heteroskedasticity. This yields the wider 95% confidence interval for the difference in mean serum LDL levels of 1.4 to 15.6 mg/dL. The p-value has increased to 0.0186, which is still statistically significant at the 0.05 significance level. This means that we still have evidence of an association between mean serum LDL level and 5 year survival status, though our test results are now more conservative.

1. Perform statistical analyses evaluating an association between serum LDL and 5 year all-cause mortality by comparing mean LDL values across groups defined by vital status at 5 years using a linear regression model that allows for the possibility of unequal variances across groups. How do the results of this analysis differ from those in problem 3? (Again, we do not need a formal report of the inference.)

**Answer**: When we fit a linear regression model without assuming equal variances across groups, the estimated difference in mean serum LDL levels remains 8.5 mg/dL. The estimate for the standard error is still 3.6 mg/dL, yielding an estimated 95% confidence interval for the true difference in mean LDL values from 1.5 to 15.5 mg/dL, which is slightly narrower than the interval in problem 3. The p-value for the test that the difference in mean LDL differs significantly from zero is now 0.017, which is slightly lower than the p-value found in problem 3. The result of this analysis is still that we have statistically significant evidence at the 0.05 significance level that the true difference in mean serum LDL levels is not zero.

1. Perform a regression analysis evaluating an association between serum LDL and age by comparing the distribution of LDL across groups defined by age as a continuous variable. (Provide formal inference where asked to.)
	1. Provide descriptive statistics appropriate to the question of an association between LDL and age. Include descriptive statistics that would help evaluate whether any such association might be confounded or modified by sex. (But we do not consider sex in the later parts of this problem.)

**Answer**: We stratified the 725 individuals with serum LDL measurements by both gender and by age in three intervals (60-69, 70-79, 80+ years old). Among all individuals in the 60-69 years age range, mean LDL was 127.7 mg/dL. The mean was observed to be lower in 70-79 year olds, who had mean LDL of 125.9 md/dL. Mean LDL was again lower in subjects who were at least 80 years old at 123.7 mg/dL. Within the 365 females in the study, we did not see the same decreasing trend in LDL with age. For the 360 males, we again observed a tendency for the mean LDL to decrease as age increased. Females did have a higher average LDL than males in the study, so future analyses may want to treat sex as an effect modifier.

|  |  |
| --- | --- |
|  | **Serum Low Density Lipoprotein (mg/dL) by Age** |
|  | **Ages 60-69** | **Ages 70-79** | **Ages 80-99** | **Any Age** |
| **Either Gender** | 127.7; n=114(32.4; 51-217) | 125.9; n=487(33.6; 11-247) |  123.7; n=124(34.7; 52-227) | 125.8; n=725(33.6; 11-247) |
| **Female** | 127.0; n=60 (34.0; 51-217) | 131.9; n=245(34.6; 11-247) | 131.1; n=60(33.5; 68-216) | 130.9; n=365(34.3; 11-247) |
| **Male** | 128.5; n=54(30.8; 68-206) | 119.8; n=242(31.6; 37-218) | 116.8; n=64(34.7; 52-227) | 120.6; n=360(32.1; 37-227) |
| *Descriptive statistics are mean; sample size (standard deviation; minimum-maximum) for LDL* |

* 1. Provide a description of the statistical methods for the model you fit to address the question of an association between LDL and age.

**Answer**: We fit a simple regression model with age in years as the sole predictor of serum LDL, measured in mg/dL. We assumed that groups across age had equal variance for this model.

* 1. Is this a saturated model? Explain your answer.

**Answer**: The model is not saturated because the number of groups does not equal the number of parameters to be estimated. There are two model parameters, an intercept and a slope, but an uncountable number of groups since the groups are being defined by age as a continuous variable.

* 1. Based on your regression model, what is the estimated mean LDL level among a population of 70 year old subjects?

**Answer**: Based on the regression model, the estimated mean LDL level among a population of 70 year old individuals is 132.53 + (-.09) \* 70 = 126.21 mg/dL.

* 1. Based on your regression model, what is the estimated mean LDL level among a population of 71 year old subjects? How does the difference between your answer to this problem and your answer to part c relate to the slope?

**Answer**: Based on the regression model, the estimated mean LDL level among a population of 71 year old individuals is 132.53 + (-.09) \* 71 = 126.12 mg/dL. Taking the difference of these predictions that are one year apart we obtain the estimated slope from our model, 126.12 - 126.21 = -.09.

* 1. Based on your regression model, what is the estimated mean LDL level among a population of 75 year old subjects? How does the difference between your answer to this problem and your answer to part c relate to the slope?

**Answer**: Based on the regression model, the estimated mean LDL level among a population of 75 year old individuals is 132.53 + (-.09) \* 75 = 125.76 mg/dL. Taking the difference between this estimate for 75 year olds and our prior estimate for 70 year olds we get five times the slope estimate, 125.76 - 126.21 = -.45 = -.09\*5.

* 1. What is the interpretation of the “root mean squared error” in your regression model?

**Answer**: The root mean squared error of 33.6 mg/dL is a pooled estimate of the sample standard deviation for serum LDL across all age groups.

* 1. What is the interpretation of the intercept? Does it have a relevant scientific interpretation?

**Answer**: The intercept represents the estimated mean serum LDL level for a population of newborns. It is scientifically unlikely that the average newborn has serum LDL of 132.5 mg/dL, which is “borderline high” according to the Mayo Clinic recommendations. The subjects in the study at hand ranged from 65 to 99 years old, so estimating newborn serum LDL based on this model makes no sense.

* 1. What is the interpretation of the slope?

**Answer**: If we compare two individuals who differ in age by one year, the slope estimate predicts that on average, the serum LDL level will be .09 mg/dL lower in the older individual.

* 1. Provide full statistical inference about an association between serum LDL and age based on your regression model.

**Answer**: The best linear fit for serum LDL in mg/dL and age estimates that, for each year two individuals differ in age, the older individual will have a serum LDL of .09 mg/dL lower per year on average, with a standard error of 0.23 mg/dL per year. Relative to a 95% confidence interval, this difference is consistent with a true difference in serum LDL of 0.54 mg/dL per year lower to 0.36 mg/dL per year higher. Using the t-test which assumes equal variances across ages, the observed rate of change in serum LDL is not significant at the 0.05 significance level (two-sided p-value = 0.694). We cannot reject the null hypothesis that age is not associated with serum LDL.

* 1. Suppose we wanted an estimate and CI for the difference in mean LDL across groups that differ by 5 years in age. What would you report?

**Answer**: We previously calculated the mean LDL across groups differing in age by 5 years to be -0.45 mg/dL. Based on a 95% confidence interval, this difference would be consistent with a true mean difference in LDL of -2.70 mg/dL to 1.80 mg/dL.

We can recalculate these estimated differences by multiplying the slope estimate and its 95% confidence interval from our regression model by 5 years since the slope is in terms of md/dL per year.

* 1. Perform a test for a nonzero correlation between LDL and age. How does your regression-based conclusion about an association between LDL and age compare to inference about correlation?

**Answer**: The sample correlation coefficient between serum LDL and age is -0.01. The test for non-zero correlation between LDL has a p-value of 0.6944, which is identical to the p-value for the regression-based test. Thus, we cannot reject the null hypothesis that serum LDL and age are uncorrelated. We have again failed to find evidence of an association between serum LDL and age.