**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2014

**Homework #2**

January 13, 2014

**Written problems:** To be submitted as a MS-Word compatible file to the class Catalyst dropbox by 9:30 am on Tuesday, January 21, 2014. See the instructions for peer grading of the homework that are posted on the web pages.

*On this (as all homeworks) Stata / R code and unedited Stata / R output is* ***TOTALLY*** *unacceptable. Instead, prepare a table of statistics gleaned from the Stata output. The table should be appropriate for inclusion in a scientific report, with all statistics rounded to a reasonable number of significant digits. (I am interested in how statistics are used to answer the scientific question.)*

***Unless explicitly told otherwise in the statement of the problem, in all problems requesting “statistical analyses” (either descriptive or inferential), you should present both***

* ***Methods: A brief sentence or paragraph describing the statistical methods you used. This should be using wording suitable for a scientific journal, though it might be a little more detailed. A reader should be able to reproduce your analysis. DO NOT PROVIDE Stata OR R CODE.***
* ***Inference: A paragraph providing full statistical inference in answer to the question. Please see the supplementary document relating to “Reporting Associations” for details.***

This homework builds on the analyses performed in homework #1, As such, 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?

The total sample size is 735. However, since 10 people have missing value on LDL level, these 10 people are omitted from all the analyses. Among the remaining 725 people, 119 died within 5 years and 606 people survived beyond 5 years. Among the 119 people who died within 5 years, the mean serum LDL level is 118.7 mg/dL with standard deviation of 36.2 mg/dL. Among the 606 people who survived beyond 5 years, the mean serum LDL level is 127.2 mg/dL with standard deviation of 32.9 mg/dL. It seems that both the sample mean and the sample standard deviation of the two groups differ significantly in magnitude.

* 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.

**Method**: Using the sample mean of LDL level to estimate the true mean of LDL level in each group. The standard error for the estimate of the true mean is sample standard deviation/sqrt(sample size). And the 95% confidence interval for estimated mean is estimated using the formula estimated mean+/- 1.96\*standard Error.

**Inference**: In the group surviving at least 5 years, the point estimate for the true mean LDLlevel is 127.2 mg/dL (95% CI:[124.6, 129.8]) and the standard error for the mean LDL level is 1.34 mg/dL. In the group having died within 5 years, the point estimate for the true mean LDLlevel is 118.7mg/dL (95% CI:[112.1, 125.3]) and the standard error for the mean LDL level is 3.31 mg/dL. It seems that both point estimates and standard errors of the two groups are very different. Since I use the sample means to estimate the true means, the point estimates are the same with the sample means. The standard errors and standard deviations are different because se=sd/sqrt(n). The deviation for the sample mean (standard error) is much smaller than the deviation for the sample.

* 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?

The CIs for the mean LDL in the two groups do overlap but we cannot draw any conclusions on the significance of the estimated difference in the estimated means basing on this fact.

* 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?)

**Method:** If we assume that the variances are equal and the means are not in the two populations, we used a pooled estimate of the variance for the combined group. The formula is , where Sx2 and Sy2 are the sample variances for the two groups respectively.

**Inference:** Using the pool method, the best estimate for the standard deviation of LDL measurements in each group is 33.48 mg/dL

* 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?

**Method:** I will use T-test assuming equal variance to compare the means LDL in the two groups and test the difference in the means.

**Inference:** Using a T-test assuming equal variance, the point estimate for the true difference in means between the two groups is 8.5 mg/dL (95% CI:[1.9, 15.1]) and the standard error of the point estimate is 3.36 mg/dL, The P value testing the hypothesis that the two populations have the same mean LDL is 0.012 indicating that the difference in the mean LDL level of the groups is significant. Therefore, I can draw conclusion that there is a statistically significant association between serum LDL and 5 year all cause mortality in this dataset.

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.

For model A, the indicator mort equals 1 if the subject died within 5 years (ldl=b0+b1\*mort). For model B the indicator suve equals 1 if the subject survived at least 5 years (ldl=b0+b1\*surv). In both models, the number of groups (2) equals to the number of parameters (2) in the model. Therefore, both models are saturated model.

* 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?

I used the model A using the indicator “mort” to indicate death within 5 years. Using this model, the estimate of the true mean LDL among a population of subjects who survive at least 5 years is 127.2mg/dL (95% CI: [124.5, 129.9]). The estimate is exactly the same as the 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 survive at least 5 years? How does this compare to the corresponding estimate from problem 1? Explain the source of any differences.

I used the model A using the indicator “mort” to indicate death within 5 years. Using this model, the 95% confidence interval for the estimate of the true mean LDL among a population of subjects who survive at least 5 years is [124.528, 129.868]. This estimate is a little bit different compared with the estimate from problem 1 [124.571, 129.825] because in the regression model we use the pooled SD to estimate the standard error and the 95% confidence interval.

* 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?

I used the model B using the indicator “surv” to indicate survival beyond 5 years. Using this model, the estimate of the true mean LDL among a population of subjects who died within 5 years is 118.7mg/dL (95% CI: [112.7, 124.7]). The estimate is exactly the same as the 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.

I used the model B using the indicator “surv” to indicate survival beyond 5 years. Using this model, the 95% confidence interval for the estimate of the true mean LDL among a population of subjects who died within 5 years is [112.673, 124.722](mg/dL). This estimate is a little bit different compared with the estimate from problem 1 [112.134, 125.261] (mg/dL) because in the regression model we use the pooled SD to estimate the standard error and the 95% confidence interval.

* 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?

If we assume the variances are equal in the two populations, the regression based estimate of the standard deviation within each group is 33.48 (Root MSE) for both models. This pool estimate for the standard deviation is exactly the same as the estimate from problem 1.

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

The reference group in model A and B is different, but the estimate of mean level of LDL in each group and the estimate of the difference mean LDL level in each group are the same for both models.

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

In model A the intercept is 127.2 meaning that the mean serum LDL level among a population of subjects who survive at least 5 years is 127.2 mg/dL.

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

The intercept in model A is -8.5, meaning that the mean serum LDL level among a population of subjects who died within 5 years is 8.5 mg/dL lower than the mean serum LDL level among a population of subjects who survived beyond 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?

Using the model A, the point estimate, the estimated standard error of the point estimate and the 95% confidence interval for the true difference in means between the two populations are 8.5 mg/dL, 3.36 mg/dL and [1.9, 15.1] respectively. The two-sided P value testing the hypothesis that the two populations have the same mean LDL is 0.012. Since the p-value is smaller than 0.05, I conclude that there is a statistically significant association between serum LDL and 5 year all-cause mortality. These inferences are exactly the same as the ones made from problem 1 using t-test assuming equal variances.

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.)

Using a t-test allowing unequal variances, the point estimate for the difference in means of the two groups is 8.5 mg/dL (95% CI: [1.4, 15.6]). The two-sided p-value is 0.019. The point estimate is the same as the one estimated assuming equal variance. However, the standard error is larger, the 95% confidence interval is broader and the p-value is bigger than the estimates from t-test assuming equal variance.

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.)

Using a linear regression model allowing heteroscedasticity, the point estimate for the difference in means of the two groups is 8.5 mg/dL (95% CI: [1.5, 15.5]). The two-sided p-value is 0.017. The point estimate is the same as the one estimated in problem 3. However, the standard error, the 95% confidence interval and the p-value are all different from the estimates using t-test allowing unequal variances.

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.)

**Table 1: Participants’ characteristics by LDL serum level**

|  |  |  |
| --- | --- | --- |
|  | Male (N=360) | Female (N=365) |
| Age groups | 65- 74 (n=204; 56.7%) | 75+ (n=156; 43.3%) | 65-75 (n=213; 58.4%) | 75+ (n=152; 41.6%) |
| Serum LDL (mg/dL) | 122.0 (31.4; 37-206) | 118.8 (33.1; 39-227) | 129.8 (33.1; 46-247) | 132.5 (35.9; 11-225) |

**Descriptive statistics presented are the mean (standard deviation; minimum – maximum)**

From the table above, we can see that the serum LDL and the age seems to be associated and the effect of age on serum LDL may be modified by sex as well. However, since sex and age do not seem to be associated, sex may not be a confounder.

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

To examine the association between LDL and age, I would fit a linear regression allowing heteroscedasticity. The serum level of LDL and age are the response and the predictor of interest respectively (ldl= b0+ b1\*age).

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

No, this is not a saturated model because the number of parameters does not equal to the number of groups (Since age is a continuous variable, there are infinite groups to compare).

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

Based on my model, the estimated mean LDL level among a population of 70 year old subjects is 126.2 mg/dL (95% CI: [123.1, 129.4])

* 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 d relate to the slope?

The estimated mean LDL level among a population of 71 year old subjects is 126.1 mg/dL (95% CI: [123.2, 129.0]). The difference between the answers of part e and part d (-0.09 mg/dL) is the slope of the regression line.

* 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?

The estimated mean LDL level among a population of 75 year old subjects is 125.8 mg/dL (95% CI: [123.3, 128.2]). The difference between the answers of part f and part d (-0.45 mg/dL) is five times of the slope of the regression line

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

The root mean squared error in my regression model is the weighted average of the standard deviations across all the age groups in the sample (pooled SD). The pooled standard deviation of LDL level is 33.6 mg/dL.

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

The intercept represents the mean serum LDL level among a population of 0 year old subjects. It does not have a relevant scientific interpretation.

* 1. What is the interpretation of the slope?

The slope represents the marginal effect of age on the mean serum LDL level. In my regression model, the slope is -0.09 meaning that with each additional year increase in the age, the mean LDL level decreases by 0.09 mg/dL.

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

In my regression model, with each unit increase in age, the serum LDL level decreases by 0.09 mg/dL (95% CI: [-0.36, 0.54]). The two-sided p-value testing the slope being 0 is 0.694 suggesting that we cannot with high confidence reject the null hypothesis that age is not associated with serum LDL levels.

* 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?

Using my linear regression model, I would report the 5 times of the slope of the estimated line. In my model, the difference in mean LDL across groups that differ by 5 years in age is 0.45 mg/dL (95% CI: [-1.83, 2.73]).

* 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?

**Method:** I would use Pearson correlation to test for a nonzero correlation between LDL and age.

**Inference:** The p-value for the correlation test is 0.694 suggesting that we cannot with high confidence reject the null hypothesis that age is not associated with serum LDL level, which is the same with the regression-based conclusion.

**Discussion Sections: January 13 – 17, 2014**

We will discuss the dataset regarding FEV and smoking in children. Come do discussion section prepared to describe the approach to the scientific question posed in the documentation file fev.doc.