**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 result is the following. The sample size for subjects who survived at least 5 yeas is 606, the sample mean is 127.2 mg/dL, and standard deviation is 32.9 mg/dL. For subjects who died in 5 years, the sample size is 119, sample mean is 118.7 mg/dL, and standard deviation is 36.2 mg/dL. The sample means and standard deviation are different between two groups. It is slightly higher sample mean but lower standard deviation in the groups with subjects survived at least 5 years.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mean | N | SD |
| Survived at least 5 years | 606 | 127.2 | 32.9 |
| Died in 5 years | 119 | 118.7 | 36.2 |

* 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.
* The point estimate of mean LDL for those who survived at least 5 years is 127.2 mg/dL, its standard error is 1.3 mg/dL, and 95% CI: (124.6, 129.8). With 95% confidence, it is not unusual if the true mean LDL is between 124.6 mg/dL and 129.8 mg/dL in the survived at least 5 years group. For those who died within 5 years, the estimated mean LDL is 118.7 mg/dL, standard error is 3.3 mg/dL, and 95% CI: (112.1, 125.3). Based on this 95% confidence interval, we can say it is not surprised if the true mean LDL is between 112.1 mg/dL and 125.3 mg/dL.
* The difference of the point estimate between two groups is as same as the difference of the sample mean, but it is totally difference between the sample standard deviation and estimated standard error. The difference of SE between two groups is larger than SD because of$ SE=\frac{SD}{\sqrt{n}}$ . Because the sample size is smaller in died in 5 years group to result in the larger difference of SE between two groups.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | N | Mean | SE | 95% CI |
| Survived at least 5 years | 606 | 127.2 | 1.3 | (124.6, 129.8) |
| Died in 5 years | 119 | 118.7 | 3.3 | (112.1, 125.3) |

* 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?
* Yes, CI is overlap. Maybe there is no enough evidence to show the mean LDL between two groups is different when α= 0.05. Further test is needed.
	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?)
* Use pool SD. Weight SD depend on sample size.
	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?
* When we use two sample, two-sided and equal variance t-test to compare the mean different LDL between two groups, we get the point estimate is 8.5 mg/dL, SE is 3.4 mg/dL and 95% CI: (1.9, 15.1). That is, we are not surprised if the true mean difference of LDL between 1.9 mg/dL and 15.1 mg/dL.

H0: the difference of mean LDL between two groups = 0

H1: the difference of mean LDL between two groups $\ne $ 0

* The p-value = 0.01 < 0.05, it is significant to reject the null hypothesis. That is, the mean LDL between two groups is different.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mean | SE | 95% CI |
| Mean Difference | 8.5 | 3.4 | (1.9, 15.1) |

 t = 2.5, p-value= 0.01

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.
* Model A: died within 5 years=1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | β | SE  | t | p-value | 95% CI. |
| Survival status  | -8.5 | 3.4 | -2.53 | 0.012 | (-15.1, -1.9) |
| Intercept | 127.2 | 1.4 | 93.53 | 0 | (124.5, 129.9) |

* Model B: died within 5 years=0

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | β | SE  | t | p-value | 95% CI. |
| Survival status | 8.5 | 3.4 | 2.53 | 0.012 | (1.9, 15.1) |
| Intercept | 118.7 | 3.1 | 38.68 | 0 | (112.7, 124.7) |

* Both are saturated model. There are only two values for predictor (died or survived), and only two parameters (intercept & beta) in model A and B.
	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?
* In model A, we know the mean LDL for subjects who survived after 5 years is 127.2 mg/dL. We can get the same result from model B as well. The result is as same as the result 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.
* From model A, we can see the 95% CI for survived after 5 years group is (124.5, 129.9) mg/dL. The 95% CI from problem 1 is (124.6, 129.8), it is only very slightly difference because of pooled SD is used in regression 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 die within 5 years? How does this compare to the corresponding estimate from problem 1?
* The mean LDL among those who died within 5 years is 118.7 mg/dL. As mentioned in 2b, the result here is as same as 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 die within 5 years? How does this compare to the corresponding estimate from problem 1? Explain the source of any differences.
* From model B, we know the 95% CI for died after 5 years group is (112.7, 124.7) . The 95% CI in problem 1 for died after 5 years is (112.1, 125.3), there is slightly difference because the regression model used pooled SD.
	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?
* Root MSE
	1. How do models A and B relate to each other?

Model A: $Y\_{i}= 127.2-8.5×survival status\_{i}$ , where died within 5 years =1

Model B: $ Y\_{i}= 118.7+8.5×survival status\_{i}$ , where died within 5 years =0

* The intercept in model A is equal to the intercept + slope in model B.
	1. Provide an interpretation of the intercept from the regression model A.
* The intercept is used when predictor=0, so we can see the mean LDL for subjects who survived after 5 years is 127.2 mg/dL from in model A.
	1. Provide an interpretation of the slope from the regression model A.
* In model A, for those people who died within 5 years, their mean LDL is 8.5 mg/dL lower than those who survived after 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?
* The result is the following. By using model A, the estimator for intercept is 127.2, for slope is -8.5, 95% CI for mean difference: (-15.1, -1.9). With 95% confidence, we are not surprised if the true mean difference of LDL for those who died within 5 years is between 1.9 mg/dL and 15.1 mg/dL lower than those who survived after 5 years.

H0: there is no linear association between LDL and mortality

H1: LDL and 5 year mortality has linear association

We have p-value =0.01, it is significant to reject the null hypothesis at alpha 0.05 level. There is linear association between LDL and mortality. The result is the same as problem 1.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | β | SE  | t | p-value | 95% CI. |
| Survival status  | -8.5 | 3.4 | -2.53 | 0.012 | (-15.1, -1.9) |
| Intercept | 127.2 | 1.4 | 93.53 | 0 | (124.5, 129.9) |

 F(1,723) = 6.41, p-value = 0.0115

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.)
* The standard error is different from problem 1, it is slightly larger, and 95% CI is wider as well.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mean | SE | 95% CI |
| Mean Difference | 8.5 | 3.6 | (1.4, 15.6) |

 t = 2.38, p-value = 0.019

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.)
* The 95% CI is different.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | β | SE  | t | p-value | 95% CI. |
| Survival status  | -8.5 | 3.6 | -2.38 | 0.017 | (-15.5, -1.5) |
| Intercept | 127.2 | 1.3 | 95.04 | 0 | (124.6, 129.8) |

 F(1,723) = 5.68, p-value = 0.017

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.)
2. 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.)
* There are 9 missing data of LDL in male and 1 missing data in female. Gender would be a modifier because mean LDL among men is 120.6 mg/dL and mean LDL among women is 130.9 mg/dL.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Male (N=369) | Female (N=366) | Total (N=735) |
|  | Mean$ \pm $SD (Min, Max) |
| LDL (mg/dL) | 120.6$ \pm $32.1 (37, 227) | 130.9 $\pm $ 34.3 (11, 247) | 125.8 $\pm $ 33.6 (11, 247) |
| Age | 74.7 $\pm $5.6 (66, 99) | 74.4 $\pm $5.3 (65,91) | 74.6 $\pm $ 5.45 (65, 99) |



1. Provide a description of the statistical methods for the model you fit to address the question of an association between LDL and age.
* Seeing that both LDL and age are continuous variables, we should use linear regression to test if there is an association between these two variables.
1. Is this a saturated model? Explain your answer.
* No. Because age is a continuous variable, the value of age is more than two. ( We have only two parameters in this model)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | β | SE  | t | p-value | 95% CI. |
| age | -0.09 | 0.23 | -0.39 | 0.69 | (-0.54, 0.36) |
| Intercept | 132.53 | 17.15 | 7.73 | 0.00 | (98.85, 166.21) |

1. Based on your regression model, what is the estimated mean LDL level among a population of 70 year old subjects?
* The regression model: $E(LDL\_{i}|age\_{i})=132.53-0.09×age\_{i}$

Let age=70, then mean LDL level of 70 year-old population is $132.53-0.09×70=126.23$ 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?
* Let age=71, then mean LDL level of 71 year-old population is $132.53-0.09×71=123.14$ mg/dL. When we increase 1 unit for age, the estimated mean LDL will decrease by 0.09 mg/dL.
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?
* Let age=75, then estimated mean LDL of 75 year-old population is $132.53-0.09×75=125.$78 mg/dL. The difference is there is many possible values for depend on different age.
1. What is the interpretation of the “root mean squared error” in your regression model?
* RMSE is the root of the expected value of the square of the difference between estimator and true value. $\sqrt{E(true value-estimator)^{2}}$
1. What is the interpretation of the intercept? Does it have a relevant scientific interpretation?
* The intercept in this model is only for mathematically fitting our data. We do not have age=0 in our sample.
1. What is the interpretation of the slope?
* The slope in regression model means how many the mean of LDL will be changed when the age change by 1 unit.
1. Provide full statistical inference about an association between serum LDL and age based on your regression model.

H0: there is no linear association between LDL and age

H1: linear association exists between LDL and age

|  |  |  |  |
| --- | --- | --- | --- |
| Source | SS | df | MS |
| Model | 175 | 1 | 175 |
| Residual | 817288 | 723 | 1130 |
| Total | 817463 | 724 | 1129 |
| F(1, 723) = 0.15 | p-value = 0.6944 |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | β | SE  | t | p-value | 95% CI. |
| age | -0.09 | 0.23 | -0.39 | 0.694 | (-0.54, 0.36) |
| Intercept | 132.53 | 17.15 | 7.73 | 0 | (98.85, 166.21) |

* The model is $E(LDL\_{i}|Age\_{i})=132.53-0.09×Age\_{i}$

When age increases 1 unit, the mean LDL will decrease by 0.09 mg/dL. The p-value is 0.69> 0.05, fail to reject the null hypothesis. We can say there is no enough evidence to show the association between LDL and age.

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?
* We know the estimate of mean difference of LDL for every 5 years old will decrease by 0.09\*5=0.45 mg/dL because of the negative slope.
* Seeing that 95% confidence interval for slope in the regression model is for the different of mean when there is 1 unit changed of the independent variable, we can calculate 95% CI for difference 5 years in age by this way: (-0.54\*5, 0.36\*5) =(-2.7, 1.8). We can say that it is not surprised if the true mean difference of LDL for every 5 years old is between 2.7 mg/dL lower or 1.8 mg/dL higher.
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?
* We can see the correlation between LDL and age is -0.0146, and p-value is 0.6944 that is the same as the p-value in the regression model.

|  |  |  |
| --- | --- | --- |
|  | ldl | age |
| ldl | 1 |  |
| age | -0.0146 | 1 |
|  | 0.6944 |  |

**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.