**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.

**Methods: In order to identify an association between serum LDL and 5 year all-cause mortality sample observations were divided in two groups by their survival time around the 5 year mark of the study. Groups were created as follows: survived >5 years and survived </= 5 years (i.e. died within 5 years). Descriptive statistics describing the mean and standard deviation serum LDL of the two groups were calculated and are presented below. A 2-sample, 2-sided t-test assuming equal variance was used to test the null that the mean LDL of the two groups are equal.**

* 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?

|  |  |  |
| --- | --- | --- |
|  | **Mean LDL (mg/dL)** | **Standard deviation (mg/dL)** |
| **Survived > 5 years (n=606)** | **127.20** | **32.93** |
| **Died within 5 years (n=119)** | **118.70** | **36.16** |

**There are 606 observations that survived greater than 5 years. The mean LDL of this group is 127.20 mg/dL +/- 32.93 mg/dL. In contrast there were only 119 that died within 5 years. This group had a mean LDL of 118. 70 mg/dL +/- 36.16 mg/dL. The means and standard deviations for the two group are very similar, varying only by approximately 3 mg/dL. From a clinical standpoint these mean LDLs would still remain in the same category of “near ideal” LDL.**

* 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.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Mean LDL (mg/dL)** | **Standard error (mg/dL)** | **95% CI (mg/dL) [LL,UL]** |
| **Survived > 5 years (n=606)** | **127.20** | **1.34** | **[ 124.57, 129.83 ]** |
| **Survived </= 5 years (n=119)** | **118.70** | **3.31** | **[ 112.13, 125.26 ]** |

**The data show that those surviving greater than 5 years had a higher LDL than those that did not survive. The mean LDL among those surviving greater than 5 years was 127.20 mg/dL (CI: 124.57, 129.83) and 118.70 (CI: 112.13, 125.26) among those that died within 5 years. The SE and resulting confidence interval was smaller for those surviving longer, with the standard error for those who died within 5 years being more than twice that of the group that survived greater than 5 years**

* 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 95% confidence intervals about the mean LDL for the two groups overlap slightly; however, we know that confidence intervals depend on sample size, with groups with larger sample sizes having a narrower CI as see here. Thus conclusions about the statistically significance should not be reached based on this finding alone.**

* 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?)

**When testing for a difference in means between two sample populations with presumed equal variance the standard deviation of the difference of the two groups is equal to the grand or pooled standard deviation which is the square root of the average variances of the two groups:**

* 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?

**Methods: A 2-sample 2-sided t-test allowing for equal variance tested the null hypothesis that the mean difference LDL between the two groups is equal to zero at the 0.05 significance level.**

**Results:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Mean LDL (mg/dL)** | **Standard error (mg/dL)** | **95% CI (mg/dL) [LL,UL]** |
| **Survived > 5 years (n=606)** | **127.20** | **1.34** | **[ 124.57, 129.83 ]** |
| **Survived </= 5 years (n=119)** | **118.70** | **3.31** | **[ 112.13, 125.26 ]** |
| **Difference** | **8.50** | **3.36** | **[1.91, 15.09]** |

**The 2-sample, 2-sided t-test allowing for equal variance suggests that the difference of means between the groups is 8.50 (1.91-15.09) p=0.01. Thus at the 0.05 significance level, we can reject the null that the means are equal (the mean difference=0) and accept the alternative that the means are not equal.**

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.

**Methods: Two simple linear regression models were constructed to determine the relationship of mean LDL with survival status for the two groups survived > 5 years and died within 5 years. The response variable in both models is mean serum LDL in mg/dL. The two models, model the opposite outcome. Model A models the association of having died within 5 years and LDL while model B models the association between LDL and having survived more than 5 years. Estimated parameters for the two models are used to predict mean serum LDL of populations that either died within or survived beyond the 5 year timeframe of interest. Resulting mean serum LDL levels are then compared to results from the two-sample, two-sided t-test analysis assessing the difference mean serum LDL performed previously. A 0.05 significance level was employed for this analysis to determine which parameters were to remain in the model.**

* 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.

**Methods: Two simple linear regression models were constructed to determine the relationship of mean LDL with survival status for the two groups survived > 5 years and survived </= 5 years. The response variable in both models is mean LDL. Model A models the association of having died within 5 years and LDL while model B models the association between LDL and having survived more than 5 years. The resulting models from this analysis are as show below**

**Model A (n=): LDL= 127.20-8.501died**

**Model B (n=): LDL= 8.501survived +118.70**

**Model A can be interpreted as among those having died within 5 years the mean LDL is equivalent to the 118.70 mg/dL or (127.20-8.5)mg/dL. Whereas model B can be interpreted as those having survived greater than 5 years of the study will have a mean LDL equal to 127.20mg/dL. These two models are saturated, because the predictor of interest in binary and the regression analysis yields two parameters with which to predict the response variable, thus the model will fit the data perfectly.**

* 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?

**Model B will be used to show the estimate for the true mean LDL among a population of subjects who survive at least 5 years :**

**Model B: LDL=8.501survived + 118.70 🡺 LDL= 8.501(1)+118.70 =127.20 mg /dL**

**The resulting value of 127.20 mg/ dL is equal to the calculated mean among those who survived greater than 5 years completed in problem one.**

* 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.

**Model A (which modeled mean LDL of those who died within 5 years of the studay) was used to determine the confidence interval for the true mean LDL among survivors. The confidence interval for the true mean LDL among a population of subjects who survived at least 5 years is 124.53, 129.87. This confidence interval is equivalent to the confidence interval created when conducting the t-test.**

* 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?

**Model A will be used to show the estimate for the true mean LDL among a population of subjects who died within 5 years :**

**Model A: LDL= 127.20-8.501died 🡺 LDL= 127.20-8.501(1) =118.70 ug /dL**

**The resulting value is equal to the calculated mean among those who died within 5 years completed 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.

**Model B (which modeled the mean LDL of those who survived > 5 years) was used to determine the confidence interval of the true mean LDL among those who died within 5 years. The confidence interval for the true mean LDL among a population of subjects who died within 5 years is the 112.67, 124.72. This confidence interval is equivalent to the confidence interval created when conducting the 2-sample t-test.**

* 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 that the variances of the two populations are equal, the regression based estimate of the standard deviation within the each group for both models is 33.477 mg/dL, which is equal to the root mean squared error. This value falls between the standard deviations of 32.93 (survived >5 years) and 36.16 (died within 5 years) found previously.**

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

**Models A and B model the inverse of the other, i.e. the model the inverse outcome of the other. Thus, the slopes for both models are the same magnitude, just opposite direction. A modeled death while B modeled survival. Because this predictor was a binary variable the intercept for the models corresponded to the opposite outcome, i.e. the mean of the other group.**

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

**Because A modeled death, the intercept is the value when the person in fact survived greater than 5 years which is equal to the mean LDL of the population that survived greater than 5 years. Thus, the interpretation of model B which tested the association between survival and serum LDL, the intercept is interpreted as the mean LDL when an individual does NOT survive at least 5 years, which is equal to the mean LDL of those who died within 5 years.**

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

**The slope of the model A is the difference in serum LDL between populations that survived greater than 5 years and those that died within 5 years. That is the amount having died changes mean serum LDL.**

* 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 true difference in mean serum LDL between a population that survives at least 5 years and a population that dies within 5 years is 8.501 mg/dL (SE: 3.36, C: 1.91, 15.09). The resulting p value from this analysis is p=0.012. Thus we can reject the null that the means of the two groups are equal and accept that null that the means for these two groups are in fact different.**

**The estimates and p-values obtained from this regression analysis are the same as those obtained from the 2-sided, 2-sample 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 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.)

**Methods: A 2-sample, 2-sided t-test allowing for unequal variances was used to compare the mean LDL across groups defined by vital status at 5 years, at the 0.05 significance level. A line will then be fitted to the data using simple linear regression to determine the association between survival and mean serum LDL.**

**Results: The mean difference of LDL identified using the 2-sample, 2-sided t-test allowing for unequal variance is 8.501 mg/dL (SE: 3.57, CI: 1.44, 15.56) p=0.019. Thus we can reject the null that the two means are the same and accept the alternative that the means are in fact different. The resulting estimates and p-value are slightly different than but similar to the estimates obtained when assuming the variances were equal.**

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.)

**Methods: Simple linear regression was used to compare the mean LDL across groups defined by vital status at 5 years.**

**Results: The resulting model from the regression analysis is as follows:**

**LDL=127.20 -8.501 (died)**

**These parameter values are nearly equivalent to the estimates obtained from both the regression methods and t-test assuming equal variance. Given that the parameter associated with death within 5 years is non-zero, we can conclude that this factor is associated with serum LDL. In the case of this model, there is an inverse relationship between serum LDL and having died within 5 years, suggesting that high LDL may actually be associated with survival. This finding agrees with the literature surrounding LDL values and overall health (and survival).**

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.)

**Methods: In order to assess if there is an association between mean serum LDL and age, a simple linear regression was performed. A table summarizing the distributions of LDL by both the predictor of interest, as well as sex was constructed to determine the presence of confounding and/or effect modification if such a relationship were to occur. Confounders and effect modifiers however were not controlled for at this time. Simple linear regression allowing unequal variance was then used to fit the best line to the model. Parameters found to be significant at the 0.05 significance level remained in the final model, indicating an association between the response variable and the predictor of interest. This model was then used to determine the mean serum LDL of populations of varying ages.**

* 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: Summary of population by age**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| age (yrs) | ldl mg/dL, mean +/- sd  | % male | weight lbs, mean +/- sd | height (cm), mean +/- sd |
| 65-69 | 127.70 +/- 32.40  | 49 | 168.77 +/- 32.3 | 168.10 +/- 9.51 |
| 70-74 | 125.32 +/- 32.50 | 50 | 163.86 +/-30.25 | 163.86 +/- 30.25 |
| 75-79 | 126.85 +/- 35.5 | 50 | 157.29 +/- 30.95 | 164.77 +/- 9.67 |
| 80-4 | 122.81 +/- 33.49 | 47 | 144.85 +/- 25.75 | 163.28 +/- 9.88 |
| 85-89 | 124.97 +/-39.13 | 54 | 149.49 +/- 22.27 | 165.91 +/- 10.11 |
| 90-94 | 124.75 +/- 35.77 | 75 | 150.63 +/- 28.61 | 166.7 +/- 8.79 |
| >=95 | 132 +/- 1.41 | 100 | 127.5 +/- 16.26 | 162.05 +/- 1.48 |
| Total | 125.80 +/- 33.60 | 50 | 159.95 +/- 30.74 | 165.77 +/- 9.71 |

**Table 1 above shows mean, standard deviation, minimum, maximum and quartiles for LDL in mg/dL and percent male by age with age shown in 5 year age categories. The ages of participants ranges from 65-99, with only 2 participants falling in the greater than 95 years-old category. The majority of participates fall in the 65-80 year range with almost half of the participants being between 70 and 75 years old. Males account for about 50% of the total population in the study though a majority of participants over the age 85 are male. Evaluating the association between mean LDL and age we see that the mean LDL decreases steadily with age. The standard deviation also increases steadily; however, only up to the 80-85 age category where it reaches its peak before dropping back down to ~35 mg/dL for the 90-95 year old category. The standard deviation for the most extreme category of greater than 95 years of age is only 1.41 mg/dL, though there are only 2 individuals in this group. Height and weight are also presented as potential factors that may explain serum LDL with overweight and obese participants expected to have lower levels.**

Wrong descriptive statistics (1)

Talking about effect modification (1)

Talking about confounding(1)

Total:2

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

**Methods: A regression line was fitted to the data to assess an association between age and LDL.**

**The resulting model is as follows:**

**LDL=132.53-0.09Age**

**that is for every unit (in this case 1 year) increase in age, the LDL will decrease by 0.09 mg/dL.**

Did not mention whether using standard error or robust standard error (1)

Did not mention that age is continuous (1)

Total: 1

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

**In this regression model we evaluated an association between age and LDL. Because we modeled age as a continuous variable the model is not saturated for there are infinite values for the age parameter used to predict mean serum LDL.**

**Total: 3**

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

**LDL=132.53-0.09Age=132.53-0.09 (70)=126.23 mg/dL**

**Using this model created from the data collected, the expected mean LDL for the population of 70 year old subjects is 126.23 mg/dL**

**Total: 3**

* 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?

**LDL=132.53-0.09Age=132.53-0.09 (71)=126.14 mg/dL**

**The slope of this model tells us for every 1 year increase in age, the mean LDL will change by 0.09 mg/dL. Thus the mean LDL for the 70 and 71 year old populations differ by only 0.09 mg/dL.**

**Total: 3**

* 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?

**LDL=132.53-0.09Age=132.53-0.09 (75)=125.78 mg/dL**

**This answer is 0.45 mg/dL lower than the value found when using 70 years. (5 units different or 5 times the slope).**

**Total: 3**

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

**The root mean squared error is the standard deviation of the sample for the regression line. For this model the rmse=33.672 mg/dL**

**Wrong**

**Total: 0**

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

**In this case, because the intercept is age, the intercept would be interpreted as follows: at 0 years of ages (or among newborns) the mean LDL is equal to the intercept, 132.53 mg/dL. The resulting value is neither scientifically relevant nor accurate.**

**Total: 3**

* 1. What is the interpretation of the slope?

**The slope of this model can be interpreted as for every one year increase in age the mean LDL decreases by 0.09 mg/dL.**

**Total: 3**

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

**Using the provided data we were able to construct a robust model of the association between age and serum LDL level in mg/dL. The resulting model suggests that for every one year increase in age mean serum LDL decreases by 0.09 mg/dL. The parameter age however was not found to be statistically significant at the 0.05 level in this case (p=0.698), thus we fail to reject the null that there is an association between age and serum LDL. It is likely that age modifies the association between LDL and another predictor. Additional tests are need to identify an association between LDL and age as an effect modifier.**

Did not mention about study population (0.5)

CI (1)

Total: 1.5

* 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?

**Because we know that one year increase in age the mean LDL decreases by 0.09 mg/dL (-0.050, 0.367), the 5 year difference in age would simply be parameter estimate and CI limits multiplied by 5. This is possible because the model is multiplicative.**

 **Thus the estimate and CI for the mean difference in LDL across groups that differ by 5 years is -0.45 mg/dL (-0.250, 1.835)**

**Total: 3**

* 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?

**The R-squared for this regression model is 0.0002. This value tells us that 0.02% of the variation in LDL is explained by age. This R-squared suggests that age does not greatly influence LDL.**

Correct correlation (1)

p-value and decision (1)

Total: 1

**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.