**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?

|  |  |  |  |
| --- | --- | --- | --- |
|  | Sample size | Sample Mean of LDL (mg/dL) | Sample Standard Deviation of LDL (mg/dL ) |
| *Survived at least 5 years* | 606 | 127 | 32.9 |
| *Died within 5 years* | 119 | 119 | 36.2 |

Above table shows the sample size, sample mean and sample standard deviation of LDL values among subjects who survived at least 5 years and subjects who died within 5 years. The sample means are not similar in magnitude, for their difference is 8.50, which is over 6% of both the mean. The sample standard deviations are also not similar, we can see there is a 3.3 difference which is about 10% of two the sample standard deviations.

* 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.

|  |  |  |  |
| --- | --- | --- | --- |
|  | ***Sample Mean of LDL (mg/dL)*** | ***Sample Standard Error of LDL (mg/dL )*** | ***95% confidence interval*** |
| ***Survived at least 5 years*** | 127 | 1.34 | [125.4, 129.8] |
| ***Died within 5 years*** | 119 | 3.31 | [112.1, 124.6] |

The point estimates are the sample mean of LDL values among subjects who died within 5 years and subjects who died within 5 years. The point estimates are not similar in magnitude, for their difference is 8.50, which is over 6% of both the mean. Are the standard errors similar in magnitude? The sample standard errors are also not similar, we can see there is a 2.0 difference which is even bigger than one of the standard errors.

There is a larger magnitude of similarity for standard deviation than in the estimated SEs because the sample size in the two groups are very different (606 and 119), the smaller group has a larger sd, so after dividing the sds with the squared sample size, we can find a much smaller se for the group survived at least 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?

From the table in b. we can see the CI for the mean LDL in a population surviving 5 years does overlap with the CI for mean LDL in a population dying with 5 years. Based on the 95% confidence intervals, this observation is statistically insignificant at a 0.05 level of significance, and we can thus conclude that there isn’t sufficient evidence to say the distribution of serum LDL differs between those who do or do not have higher risk of death over a 5-year period. And there is no sufficient evidence to show that serum LDL and 5 year all cause mortality are associated.

* 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?)

The combined estimated sample standard deviation should be calculated as follows

 = 33.5 (mg/dL)

This is not the stata output of combined sd of 33.6.

* 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?

For the true difference in means between a population that survives at least 5 years and a population that dies with 5 years: the point estimate is 8.50(mg/dL), the estimated standard error of the point estimate is 3.36, the 95% confidence interval is [1.9, 15.1]. Using a t test that presumes equal variances, this observation is statistically significant at a 0.05 level of significance (two-sided P= 0.0115), we can thus 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. Serum LDL and 5 year all cause mortality are associated. *(The statistical significance could be due to different variances. From the sample descriptive statistics, we see that the group with the smaller sample size (those dying within 5 years) also has greater variability of serum LDL measurements. If that estimated difference in SD of LDL were true in the population, use of the t test that presumes equal variances is anti-conservative: the p values are too low and the CI are too wide.)*

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.

Both model A and B are saturated models, because: The predictor variable used in the analysis only had two values: 0 = survives at least 5 years, 1 = dies within 5 years. The regression model has two parameters . Thus the number of groups = number of parameters = 2. We are not borrowing information across the groups for the mean. Each group mean can be fit exactly.

* 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 model A. The estimate of the true mean LDL among a population of subjects who survive at least 5 years is 127 mg/dL which is exactly the same estimate as in question 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 model A. The confidence interval for the true mean LDL among a population of subjects who survive at least 5 years is [124.5, 129.9] mg/dL. The corresponding estimate from problem 1 is [124.6, 129.8]. The CI from regression method is a little wider then the ttest method one. The CI for the intercept is not the CI for the t test output because in regression, we use the pooled SD

 instead of . Also the degrees of freedom for the two critical value t are also different, instead of .

* 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 model B. The estimate of the true mean LDL among a population of subjects who die within 5 years is 119 mg/dL, which is exactly the same estimate as in question 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 model B. The confidence interval for the true mean LDL among a population of subjects who survive at least 5 years is [112.7, 124.7] mg/dL. The corresponding estimate from problem 1 is [112.1, 125.3]. The CI from regression method is a little narrower then the t test method one. The CI for the intercept is not the CI for the t test output because in regression, we use the pooled SD

 instead of . Also the degrees of freedom for the two critical value t are also different, instead of .

* 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 presume the variances are equal in the two populations, the regression based estimate of the standard deviation within each group for each model is both 33.5 which is equal to the corresponding estimate from problem 1

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

Model A and model B are the same except for naming different groups as 0 and 1. A named the death within 5 years group 1and B named the survival group as 1. Consequently, we get slopes (including CI) that are opposite numbers -8.5[-15.1, -1.9], 8.5[1.9, 15.1] and t values that are opposite numbers -2.53, 2.53. The slope is the estimated difference in average LDL for two groups (group1-group 0). The intercept is the mean estimate for group 0, so in model A, it is the mean LDL among subjects who survive at least 5 years and it is the mean LDL among subjects who died in 5 years for model B. The corresponding t value, se, p value and CI are also different. Everything else for the two models should be the same.

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

The intercept is the mean estimate for group 0, so in model A, it is the mean LDL among subjects who survive at least 5 years. The intercept is 127mg/dl with 95% confidence [124.5, 129.9]. Thus our observation would not be atypical if the true mean LDL level for people in the “survived 5 years” population is from 124.5 mg/dl to 129.9 mg/dl. And the two-sided p value is less than 0.001. With a 0.05 significance level, we have high confidence to reject the null hypothesis that the mean LDL level for people in the “survived 5 years” population is not zero and that the estimated mean is 127 mg/dl.

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

The slope in model A is -8.50[-15.1, -1.9]. This is the estimated difference in means LDL level for people in the “died within 5 years” population and people in the “survived 5 years” population. Thus our observation would not be atypical if the true mean LDL level for people in the “died within 5 years” population is from 1.91 mg/dl lower to 15.1 mg/dl lower than the mean LDL level for people in the “Survived 5 years” population. And the two-sided p value is 0.012. With a 0.05 significance level, we have high confidence to reject the null hypothesis that the difference in means LDL level for people in the two populations is zero.

* 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?

I used model B. 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 8.50, the estimated standard error of the point estimate is 3.36, the 95% confidence interval is [1.9, 15.1]. The two-sided P value testing the hypothesis that the two populations have the same mean LDL is 0.012.

This observation is statistically significant at a 0.05 level of significance (two-sided P= 0.012), 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. There is a statistically significant association between serum LDL and 5 year all cause mortality. We reach the same conclusion as in question 1, but with a more significant p value.

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.)

Mean serum LDL levels were compared between subjects who died within 5 years of study enrollment and those who survived at least 5 years. Differences in the mean were tested using a t test that does not presume equality of variances. 95% confidence intervals for the difference in population means were similarly based on that same handling of variances.

Mean serum LDL was 127 mg/dL among the 606 subjects who survived at least 5 years after study enrollment and 119 mg/dL among the 119 subjects who died within 5 years. Based on a 95% confidence interval computed with an allowance for unequal variances, this observed tendency of 8.50 mg/dL lower mean serum LDL among subjects dying earlier would not be judged unusual if the true difference population means were anywhere between a 1.44 mg/dL to 15.6 mg/dL lower mean LDL among subjects who die within 5 years. The estimated standard error of the point estimate is 3.57. Using a t test that similarly allows for the possibility of unequal variances, this observation is statistically significant at a 0.05 level of significance (two-sided P= 0.0186), and we can with high confidence reject the null hypothesis that the mean serum LDL levels are not different by vital status at 5 years in favor of a hypothesis that death within 5 years is associated with lower mean serum LDL.

We reach the same conclusion as in question 1, but with a slightly less significant p value and a wider more conservative 95% confidence interval.

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.)

Mean serum LDL levels were compared between subjects who died within 5 years of study enrollment and those who survived at least 5 years using a linear regression model that allows for the possibility of unequal variances across groups. 95% confidence intervals for the difference in population means were based on robust the standard error.

Based on a 95% confidence interval computed with an allowance for unequal variances, this observed tendency of 8.50 mg/dL lower mean serum LDL among subjects dying earlier would not be judged unusual if the true difference population means were anywhere between a 1.50 mg/dL to 15.5 mg/dL lower mean LDL among subjects who die within 5 years. The estimated standard error of the point estimate is 3.57. Using a t test that similarly allows for the possibility of unequal variances, this observation is statistically significant at a 0.05 level of significance (two-sided P= 0.017), and we can with high confidence reject the null hypothesis that the mean serum LDL levels are not different by vital status at 5 years in favor of a hypothesis that death within 5 years is associated with lower mean serum LDL. The t statistics and the two sided p values are also different.

We reach the same conclusion as in question 3, but with a slightly more significant p value and a narrower less conservative 95% confidence interval, due to a smaller robust standard error.

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.)

Methods: 10 subjects with missing ldl was deleted, leaving 725 subjects. For table 1b: An indicator variable was created for age (Age < 74, Age ≥ 74). Descriptive statistics are presented within groups defined by females, males, and for the entire sample. Within each group, for low-density lipoprotein (LDL), we include the number of observations, mean and standard deviation all stratified into two age groups by the age indicator variable. For table 1a: Descriptive statistics are presented within groups defined by sex. Within each group, for age, we include the number of observations, mean, standard deviation minimum and maximum.

***Table 1a*** *Age by sex*

|  |  |  |
| --- | --- | --- |
|  | **Female n=365** | **Male n=360** |
| Mean (SD) | Min | Max | Mean (SD) | Min | Max |
| **Age** (yrs) | 74.4 (5.26) | 65 | 91 | 74.7 (5.63) | 66 | 99 |

***Table 1b*** *Serum Low Density Lipoprotein (LDL) by sex and age group*

|  |  |  |  |
| --- | --- | --- | --- |
| **LDL (mg/lD)** | **Female** n, **Mean** (SD) | **Male** n, **Mean** (SD) | **All subjects**n, **Mean** (SD) |
| **Age (yrs) < 74**  | 184, **130**(34.1) | 178, **123**(31.4) | 362, **127**(32.9) |
| **Age (yrs) ≥ 74**  | 181, **132**(34.5) | 182, **118**(32.8) | 363, **125**(34.3) |
| **All ages** | 365, **131**(34.1) | 360, **121**(32.1) | 725, **126**(33.6) |

Data is available on 735 subjects, however 10 of those subjects are missing data on serum low-density lipoprotein (LDL). Those subjects are omitted from all analyses, but it should be remembered that we could not assess the impact that such omissions might have on the generalizability of our results. None of the 725 subjects were missing data on any other variables of interest for this analysis.

Of the 725 subjects with available measurements, 365 are female and 360 are male. Table 1a, 1b presents descriptive statistics within these groups. From table 1a, we see almost no difference in age across gender groups. From table 1b: Female subjects in the sample tend to have higher LDL than male: mean serum LDL was 131 mg/dL for females compared to a mean serum LDL of 121 mg/dL in male subjects. The two age groups had almost the same number of subjects, and approximately similar number of subjects of each gender. Female subjects in the sample tend to have higher LDL when older in contrast with male whom tend to have lower mean serum LDL when older. Thus we predict that sex is a potential effect modifier for the association between serum LDL and age.

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

The robust linear regression model is used to address the question of an association between LDL and age. Least square estimation is used to estimate parameters and Huber-White sandwich method is used to adjust for the possibility of unequal variances. Mean serum LDL levels were compared across groups defined by age as a continuous variable using a linear regression model that allows for the possibility of unequal variances across groups. 95% confidence intervals were based on robust the standard errors.

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

This is not a saturated model, because: The predictor variable used in the analysis is age, which is continuous not binary, there are more than two values. The regression model has two parameters . Thus the number of groups > number of parameters = 2.

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

The estimated mean LDL level of 70-year-old subjects is 126.2 mg/dL

132.5281 -0.0901904\*70=126.2148 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?

The estimated mean LDL level of 71-year-old subjects is 126.1 mg/dL

132.5281 -0.0901904\*71=126.1246 mg/dL

I think the question meant to ask the difference with part d instead of c. The difference between my answer to this problem and my answer to part c is 0.09 = the difference in mean LDL between groups differing by 1 year in age = the slope

126.1246-126.2148=-0.0902 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?

The estimated mean LDL level of 75-year-old subjects is 125.8 mg/dL

132.5281 -0.0901904\*75=125.7638 mg/dL

I think the question meant to ask the difference with part d instead of c. The difference between my answer to this problem and my answer to part c is -0.451

125.7638-126.2148=-0.451 mg/dL

This is equal to the difference in mean LDL between groups differing by 5 year in age = 5 times the slope

= 5\* (-0.0901904) = -0.451mg/dL

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

root mean squared error” in your regression model

The root mean squared error is the sample standard deviation of the residuals. In the robust model that adjusts for the possibility of unequal variances across groups, the root mean squared error is the square root of average variances across groups. In a model that presumes constant variance in age groups, “root mean squared error” is within group standard deviation. The spread of the difference between observed value and expectation is roughly described by the root mean squared error.

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

The intercept 133mg/dL is the estimated mean when age = 0. In this problem, the intercept is just a mathematical construct to fit a line over the range of our data, it does not have a relevant scientific interpretation， because we don't really have any data on newborns.

* 1. What is the interpretation of the slope?

The slope - 0.09 mg/dL is the difference in mean LDL between groups differing by 1 year in age, with the older group averaging a lower score.

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

From linear regression analysis, we estimate that for each year difference in age (one year older), the difference in mean serum low-density lipoprotein (LDL) is - 0.09 mg/dL. A 95% CI suggests that this observation is not unusual if the true difference in mean LDL per year difference in age were between -0.55 and 0.37 mg/dL. Because the P value is 0.698 > 0.05, we failed to reject the null hypothesis that there is no linear trend in the average LDL across age groups.

* 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?

The difference in mean LDL across groups that differ by 5 years in age is -0.45 mg/dL, the 95% confidence interval is [-2.73, 1.83]

* 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?

I categorized age into 7 groups, each ranging 5 years. I also categorized ldl into 5 groups according to the table in homework1. Then I used the chi-squared test on these two new variables. The p value is 0.762>>0.05, we failed to reject the null hypothesis, that there is no association in the average LDL across age groups. This is consistent 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.