**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?

ANS: For this question I consider the death within 5 years as done in the answer key for the last assignment so we are looking at time <= 5\*365.25 = 1826.25. The subjects who survived at least 5 years had a sample size of 606 with a sample mean of 127.2 mg/dL and a sample standard deviation of 32.9 mg/dL. The subjects who died within 5 years had a sample size of 119 with a sample mean of 118.7 mg/dL and a sample standard deviation of 36.2 mg/dL.

The sample mean is 8.3 mg/dL higher for those who survived atleast 5 years and the standard deviation is 3.3 mg/dL lower for those who survived atleast 5 years but based only on these sample statistics we cannot say if the difference is statistically or scientifically significant.

* 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 is equivalent to the sample mean and the standard error is the ratio of sample standard deviation divided by the square root of the sample. The 95% confidence interval is obtained as a standard WALD type statistic using a t-distribution.

For those who died within 5 years the point estimate is 118.7mg/dL with estimated standard error 3.31 mg/dL with a 95% confidence interval (112.1 mg/dL, 125.3 mg/dL). For those who died after 5 years the point estimate is 127.2 mg/dL with estimated standard error 1.34 mg/dL with a 95% confidence interval (124.6 mg/dL, 129.8 mg/dL). The point estimates differ by the same amount as the sample means and the standard errors differ by 1.97 mg/dL. The sample mean estimated population mean is the same since the sample mean is an unbiased estimator of the population mean. The difference between the standard error is less than the difference between the sample standard deviation which can be explained by the fact that we account for the sample size when evaluating the standard error. So if the two sample sizes were the same the difference between standard errors and standard deviations would be the same.

* 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 confidence intervals overlap. This implies that

* 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 best estimate we have is the pooled standard deviation given by
$$s\_{X1X2}= \sqrt{\frac{\left(n1-1\right)s\_{X1}^{2}+(n2-1)s\_{X2}^{2}}{n1+n2-2}}=33.5 mg/dL$$

* 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?

The point estimate for the difference of mean is 8.50 mg/dL higher for the 606 patients who survived at-least 5 years with and estimated standard error of 3.36 mg/dL with a 95% confidence interval of (1.91, 15.1) mg/dL higher for those who survived at-least 5 years. The P-value for a two sided t-test for independent samples is 0.012. Based on the two sided P-value 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.

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.

We have for Model A the fitted model:

E(serum ldl) = 127.2 – 8.50**\*I(**subject died within 5 years)

For Model B we have the fitted model:

E(serum ldl) = 118.7 + 8.50**\*I(**subject died after 5 years)

In both cases we have a saturated model because the predictor is binary variable which takes two different values 0 and 1 and the number of parameters of the estimated models are also two. In both cases we wish to estimate the intercept and slope.

* 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?

Using Model A the estimated mean of those who survived atleast 5 years is given when **I(**subject died within 5 years) = 0 and so the estimate is just the intercept of Model A which is 127.2 mg/dL. This is coincides with our 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.

Using Model A again, we have the confidence interval (124.5, 129.9) mg/dL which is the confidence interval for the intercept in the model. In this case we observe a slightly wider confidence interval or a more conservative confidence interval. We observe a small difference because under the assumptions of homoscedasticity the standard error used to evaluate the confidence interval is the pooled standard error which borrows information from the second group too. The difference is fairly small (0.2) because the difference in sample standard deviations/standard errors is not that large and so the pooled estimate of standard error is close to the standard error of the single group sample.

* 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?

This time using model B, the estimated mean of those who dies within 5 years is obtain when we set **I(**subject died after 5 years)=0 and so now we’re looking at the intercept of model B which is 118.7 mg/dL. This estimate coincides with the estimate obtained 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.

Once again we’re using model B and so the parameter of interest is the intercept of the model. In this case the 95% confidence interval is (112.7, 124.7) mg/dL. We observe that this confidence interval is smaller or more anti-conservative. Again, the difference can be explained by the fact for the regression estimates we used the pooled estimate.

* 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?

The estimated standard deviation of each group for model A is 33.5 mg/dL and the estimated standard error of each group in model B is also 33.5 mg/dL. This is equal to the pooled standard deviation we obtained in problem which shows that in this case the two are equivalent.

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

Models A is the fitted model for the re-parameterized predictor in model B. If the predictor in model B is X = **I(**subject died after 5 years) then consider:

E(serum ldl) = 118.7 + 8.50**\***(1-X) = 127.2-8.50\*X

Which shows the relationship between model A and B.

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

The intercept for model A is the estimated mean serum LDL level for the group of patients who did NOT die within 5 years. In other words it is the mean serum LDL level for patients who lived at-least 5 years.

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

The intercept of the model A is the estimated difference of mean serum LDL level between the group of patients who died within 5 years and those who lived for at-least 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?

In this case I will use model B because the intercept of this model is for the estimated quantity of interest. Though due to the relationship between the two model means the estimates just the negation of the other for the point estimate and confidence interval but the standard error is the same. The point estimate using the regression model B is 8.50 mg/dL higher for those who survived at-least 5 years with an estimated standard error of 3.36 mg/dL for the difference in mean serum LDL and a 95% confidence interval (1.91, 15.09) mg/dL higher for those who survived at-least 5 years. The P-value obtained in this case is 0.012 <0.05 and so with high confidence we reject null hypothesis in favor of the alternative that the mean serum LDL differs across the two groups. The p-value and confidence interval is the same as we had for problem and subsequently we have the same conclusions. In this case we use the pooled standard error for our t-test and so because of that the estimates and p-values are the same for the t-test which assumes 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.)

For this question the point estimates for the mean serum LDL will be the same as in the previous parts the only things which will differ will be the confidence intervals and standard errors for the difference in means and consequently the P-value for the test. So we perform a t-test for the difference of means allowing for un-equal variance for independent samples. In this case we observe a standard error of 3.57 and a 95% confidence interval of (1.44, 15.56) mg/dL higher for those who survived at-least 5 years with a P-value of 0.019. Based on the low P-value and the confidence interval we reject the null hypothesis in favor of the alternative that the mean serum LDL level differs across the two groups.

In this analysis we observed a wider confidence interval and higher P-value which indicates that under the assumption of allowing for unequal variance we have a more conservative result because we have a higher standard error.

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.)

For this question the point estimates for the mean serum LDL will be the same as in question 2, the only things which will differ will be the confidence intervals and standard errors for the difference in means and consequently the P-value for the test. So we perform a robust regression using the predictor the indicator variable of those who survived at-least 5 years. In this case we observe a standard error of 3.57 and a 95% confidence interval of (1.50, 15.50) mg/dL higher for those who survived at-least 5 years with a P-value of 0.017. Based on the low P-value and the confidence interval we reject the null hypothesis in favor of the alternative that the mean serum LDL level differs across the two groups.

In this analysis we observed that the standard error did not change for upto 1 decimal place which means any differences are small. We observe narrower confidence interval and a lower P-value indicative of an anti-conservative estimate compared to the t-test of question 3

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.)

The table below gives the descriptive statistics which would be appropriate for considering an association between LDL and age.

|  |  |
| --- | --- |
|  | Serum LDL (mg/dL) |
|  | N(%) | Meanmg/dL | SDmg/dL | Minmg/dL | Maxmg/dL |
| Age 65-74 years | 417 (57.5%) | 126.0 | 32.5 | 37 | 247 |
| Age 75-84 years | 264 (36.4%) | 125.6 | 34.9 | 11 | 227 |
| Age >=85 years | 44 (6.1%) | 125.2 | 37.2 | 57 | 216 |
| All Ages | 725 (100%) | 125.8 | 33.6 | 11 | 247 |

Based on the table above we observe that across the different age groups the Mean serum LDL level does not vary by a large amount and the standard deviations are fairly similar too which shows that there is no obvious association between age and serum LDL however this hypothesis still needs to be tested by a formal statistical analysis.

If we wish to consider the possible effect of SEX as a confounder or effect modifier the table of descriptive statistics is presented to consider possible associations in the sample between Sex and Age.

|  |  |
| --- | --- |
|  | Sex |
|  | No. of Males (%) | No. of Females(%) |
| Age 65-74 years | 204 (48.9%) | 213 (51.1%) |
| Age 75-84 years | 130 (49.2%) | 134 (50.8%) |
| Age >=85 years | 26 (59.1%) | 18 (40.9%) |
| All Ages | 360 (49.7%) | 365 (50.3%) |

Based on the table above we observe that the proportion of males versus females is similar for ages 65-74 and 75-84 years and higher for ages 85 and above. This may be indicative of an association but it might not be statistically significant. I would suggest performing a formal test of association for instance using robust regression.

If we wish to consider confounding or effect modification we would need to consider this causal relationship before the experiment but generally it could still be useful to verify the assumptions. The table below presents descriptive statistics which could be considered when considering the question of effect modification or confounding.

|  |  |
| --- | --- |
|  | Serum LDL (mg/dL) |
|  | N(%) | Meanmg/dL | SDmg/dL | Minmg/dL | Maxmg/dL |
| Males | 360 (49.7%) | 120.6 | 32.1 | 37.0 | 227.0 |
| Females | 365 (50.3%) | 130.9 | 34.3 | 11.0 | 247.0 |
| All patients | 725 (100%) | 125.8 | 33.6 | 11.0 | 247.0 |

From the above table we see that the mean LDL levels for males is 10.3 mg/dL lower than for females which indicates that if we assume there is a possibility of confounding then this would agree with our assumption since we observe a difference of means across the two groups. The difference may not be significant for the sample but the point estimate certainly differs across the two groups.

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

For this model we fit a robust linear regression model which does not assume homoscedasticity.

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

This is not a saturated model. A saturated model is one where the number of parameters is equal to the number of different groups in our covariates. The model only has two parameters the intercept and slope of linear regression model. For the variable age we have 31 different ages in our sample which means the number of groups = 31 is not equal to the number of parameters =2 hence we do not have a saturated model.

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

For 70 year olds we have the estimated mean LDL level as 126.2 -0.09\*70 = 126.21 mg/dL

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

For 71 year olds we have the estimated mean LDL level as 126.2 -0.09\*71 = 126.12 mg/dL. I assume the questions means part d and in this case the difference is -0.09 which is equal to the slope of the linear model.

* 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?

For 75 year olds we have the estimated mean LDL level as 126.2 -0.09\*75 = 125.76 mg/dL. I assume the questions means part d and in this case the difference is -0.49 which is equal to 5 times the slope of the linear model.

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

The root mean square error gives us the sample standard deviation of the residuals and in our since we do not assume homoscedasticity, in our model the RMSE gives us an idea of the average variance across different ages.

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

The intercept in this model represents the estimated mean of serum LDL levels for patients of age 0 which means newborns. In other words, the intercept represents the estimated mean of the serum LDL levels at birth. This could be scientifically relevant depending on the scientific question. In this case it would not be relevant because we are only concerned with elderly individuals as the youngest member of the study is 65 years old and so the intercept of this model would not be scientifically relevant.

* 1. What is the interpretation of the slope?

The slope in this case is the estimated mean of the difference of serum LDL levels between individuals whose age differ by 1 year.

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

Based on a robust regression on 725 individual were we use the variable age as a continuous predictor and the mean serum LDL level as a response. The estimated mean of the difference of serum LDL levels for in an age difference of 1 year is -0.09 mg/dL with a 95% confidence interval of 0.55 mg/dL less to 0.37 mg/dL more (-0.55 mg/dL, 0.37 mg/dL) . Which indicates that the observed results would not be surprising if the true difference in means was within the range of the confidence interval. Because the P-value is equal to 0.698 which is greater than 0.05 we fail to reject the null hypothesis and conclude that there is insufficient evidence to suggest an association between age and mean 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?

In this case I would report 5\*slope ± 1.96\*(robust SE) = (-0.91, 0.01) mg/dL

* 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 perform a Pearson’s test for correlation and be obtain a point estimate for the correlation of -0.015 with a 95% confidence interval of (-0.087, 0.058). Which indicates that the observed data would not be surprising if the true correlation was between -0.087 and 0.058. We observe P-value of 0.694 due to which we fail to reject the null hypothesis and conclude there is insufficient evidence to indicate the correlation is not zero. This test is the same as the test obtained for the significance of the covariate for the case of simple linear regression which we have in this question. Hence the P-value is the same and the test statistic reported for regression is the F statistic which is just the square of the T-statistic for the Pearson’s test

**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.